Uniwersytet Kardynała Stefana Wyszyńskiego w Warszawie

Instytut Psychologii

Tomasz Rak

Weryfikacja jednowymiarowej struktury zmiennej latentnej w konfirmacyjnej analizie czynnikowej: analiza trafności wskaźników w badaniach symulacyjnych oraz empirycznych

Praca doktorska napisana pod kierunkiem

dr hab. Jan Cieciuch, prof. ucz.

Warszawa, 2025

Spis treści

Abstrakt 5
Abstract 8
1. Wprowadzenie teoretyczne 11
1.1. Pomiar cech latentnych w psychologii 11
1.2. Konfirmacja modelu teoretycznego cechy latentnej 12
1.3. Nieprawidłowa specyfikacja modelu i jej konsekwencje 14
1.4. Metody (miary) oceny poprawności modelu teoretycznego w analizie SEM/CFA 17
1.4.1. Miary globalne 17
1.4.2. Miary przyrostowe 20
1.4.3. Indeksy modyfikacji 21
1.4.4. Konkluzje 24
1.5. Jedno- i wielowymiarowe modele teoretyczne: możliwości rozróżnienia 24
2. Problem badań własnych 27
3. Badanie 1 (dane symulacyjne) 28
3.1. Metoda badania symulacyjnego 28
3.1.1. Wprowadzenie 28
3.1.2. Przygotowanie symulacji 29
3.1.3. Kontrola warunków symulacji 31
3.1.4. Warunki brzegowe interpretacji miar dopasowania 35
3.2. Wyniki badania symulacyjnego 36
3.2.1. Wprowadzenie 36
3.2.2. Różnice współczynników w strukturach jedno- i wielowymiarowych 37
3.2.3. Użyteczność miar dopasowania w rozstrzyganiu o wymiarowości 42
3.2.3.1. Miary globalne 47
3.2.3.2. Miary przyrostowe 55
3.2.3.1. Indeksy modyfikacji 58
4. Badanie 2 (dane empiryczne) 68
4.1. Metoda badania empirycznego 68
4.2. Wyniki badania empirycznego 72
4.2.1. Użyteczność miar dopasowania w rozstrzyganiu o wymiarowości 73
4.2.1.1. Miary globalne 73
4.2.1.2. Miary przyrostowe 80
4.2.1.3. Indeksy modyfikacji 84
5. Nieefektywna droga rozróżnienia przez punkty odcięcia 91
6. Dyskusja 94
6.1 Podsumowanie i dyskusja uzyskanych wyników 94
6.2 Rekomendacje 101
6.3 Ograniczenia badań własnych oraz kierunki dalszych badań 102
Bibliografia 107
Suplement: Dodatkowe tabele i załączniki 125

ABSTRAKT

Samoopisowe kwestionariusze psychologiczne do pomiaru danej zmiennej latentnej muszą spełnić szereg kryteriów, pozwalających na ocenę tego, czy poprawnie mierzą to, do czego mierzenia zostały skonstruowane. W szczególności należy określić strukturę badanego zjawiska, czyli określić, ile ma ono wymiarów mierzonych danym narzędziem. Możliwa jest struktura jednowymiarowa (co oznacza, że dany kwestionariusz mierzy jedną zmienną latentną) lub wielowymiarowa (co oznacza, że dany kwestionariusz mierzy kilka względnie niezależnych zmiennych latentnych).

Podczas określania wymiarowości struktury, badacz może wykorzystać podejście eksploracyjne lub konfirmacyjne. W pierwszym przypadku analizy przeprowadzone na zebranych danych pozwalają określić, bez formułowania żadnych wstępnych założeń, ile i jakie wymiary są mierzone danym narzędziem; w drugim przypadku badacz przyjmuje hipotezę, ile – zgodnie z jego modelem teoretycznym – wymiarów mierzy jego narzędzie oraz które pytania kwestionariusza wchodzą w skład danego wymiaru, a następnie analizy statystyczne pozwalają potwierdzić lub odrzucić jego hipotezę. To drugie podejście jest przedmiotem zainteresowania niniejszej pracy.

W podejściu konfirmacyjnym stosowana jest konfirmacyjna analiza czynnikowa (CFA, *confirmatory factor analysis*), polegająca na weryfikacji, czy model teoretyczny jest dobrze dopasowany do danych, przy czym zwykle oceniane jest dopasowanie kilku alternatywnych modeli CFA o różnej liczbie wymiarów (zmiennych latentnych). W tym celu stosowane są tak zwane miary dopasowania, które pozwalają badaczowi podjąć decyzję, który model jest najlepiej dopasowany do danych.

Jednym z fundamentalnych rozstrzygnięć w obrębie analiz CFA jest określenie, czy dany konstrukt psychologiczny ma charakter jedno- czy wielowymiarowy, to znaczy czy jest on determinowany przez pojedynczy, spójny czynnik latentny odzwierciedlający jednorodną właściwość psychologiczną, czy też nie. Pomimo powszechnego stosowania klasycznych miar dopasowania, literatura nie dostarcza jednoznacznych odpowiedzi, czy współczynniki te rzeczywiście umożliwiają trafne rozróżnianie struktur jedno- od wielowymiarowych, zwłaszcza gdy dane zostały wygenerowane lub zebrane dla konstruktu jednoznacznie jednowymiarowego. W efekcie, badacze mogą podejmować błędne decyzje dotyczące specyfiki badanego konstruktu, co w dalszej kolejności wpływa na interpretację wyników i jakość stosowanych narzędzi diagnostycznych.

Celem niniejszej pracy jest wypełnienie tej luki i udzielenie odpowiedzi na trzy pytania badawcze: (P1) czy klasyczne współczynniki dopasowania różnicują modele jedno- i wielowymiarowe w przypadku danych jednoznacznie jednowymiarowych; (P2) które z miar dopasowania są najbardziej użyteczne przy ocenie poprawności modelu CFA; oraz (P3) czy możliwe jest określenie stabilnych, praktycznych punktów odcięcia, które umożliwiłyby trafne rozstrzyganie o wymiarowości modelu.

W celu udzielenia odpowiedzi przeprowadzono dwa badania: badanie symulacyjne (N = 31 500 zbiorów danych) oraz badanie empiryczne (N = 5172 osób badanych). W obu porównywano modele teoretyczne zgodne i niezgodne z jednowymiarową strukturą danych. Przeprowadzono analizy trafności kryteriów najczęściej stosowanych w literaturze: przetestowano globalne miary dopasowania: χ^2, χ^2/ df, χ^2_p, CFI, TLI, GFI, AGFI, RNI, RMSEA, AIC, BIC, RMR, oraz SRMR, a także mniej znane, oparte o *indeksy modyfikacji*, lokalne błędy specyfikacji: *MI*, *NCP*, *ECP*, *m*, *nm*, *EPC:m*, *EPC:nm*, oraz *i*.

W analizach symulacyjnych wykorzystano dane wygenerowane jako jednoznacznie jednowymiarowe, na których porównywano modele poprawnie określone jako jednowymiarowe oraz błędnie jako wielowymiarowe. Celem było odwzorowanie sytuacji, w której rzeczywista struktura zmiennej jest jednowymiarowa, lecz badacz błędnie zakłada większą złożoność modelu. W analizach empirycznych zastosowano podobne podejście, wykorzystując dane z rzeczywistych badań z użyciem jednowymiarowych narzędzi psychometrycznych. Tak jak w przypadku danych symulowanych, oceniano trafność klasycznych miar dopasowania i indeksów modyfikacji w odróżnianiu modeli poprawnych od błędnych, uwzględniając zakłócenia typowe dla danych empirycznych.

Wyniki uzyskany w obu badaniach (symulacyjnym oraz realnym były zgodne) i wykazały, że klasyczne współczynniki dopasowania nie pozwalają na wiarygodne rozróżnienie między modelami jedno- i wielowymiarowymi, gdy dane mają jednoznacznie jednowymiarową strukturę. Po pierwsze (P1), różnice pomiędzy modelami były nieistotne i często przypadkowe, co oznacza, że tradycyjne miary dopasowania nie są skuteczne w identyfikacji poprawnego modelu. Po drugie (P2), jedynie miary oparte na indeksach modyfikacji - w szczególności wskaźnik *nm%* – umożliwiały względnie jednoznaczną ocenę wymiarowości modelu. Po trzecie (P3), nie udało się wyznaczyć stabilnych punktów odcięcia, które pozwalałyby na automatyczne zaklasyfikowanie modelu jako jedno- lub wielowymiarowego.

Praca wskazuje na ograniczoną użyteczność klasycznych miar dopasowania w analizie modeli jednowymiarowych oraz na potencjał podejścia opartego na analizie lokalnych błędów

specyfikacji oraz proponuje jako alternatywną drogę wykorzystanie tak zwanych *indeksów modyfikacji*, czyli, w odróżnieniu od wyżej wymienionych globalnych miar dopasowania, miar wykorzystywanych obecnie w określaniu poprawności specyfikacji poszczególnych obiektów w modelu (osobnych ścieżek, czy „połączeń" konkretnych pytań kwestionariusza w czynniki). Przeprowadzone analizy pozwoliły sformułować rekomendacje dotyczące stosowania alternatywnego podejścia z użyciem tzw. indeksów modyfikacji, które wykazują spory potencjał i pozwalają na dokonanie poprawnych rozróżnień tam, gdzie jest tylko jeden silny czynnik i występują wątpliwości co do struktury operacjonalizowanego konstruktu.

Słowa kluczowe: konfirmacyjna analiza czynnikowa, struktura, symulacje, współczynniki dopasowania, zmienna latentna

ABSTRACT

Self-report psychological questionnaires used to measure a latent variable must meet a number of criteria that allow researchers to assess whether they accurately measure the construct for which they were developed. In particular, it is necessary to define the structure of the studied phenomenon, that is, to determine how many dimensions are measured by a given instrument. The structure may be unidimensional (meaning that the questionnaire measures a single latent variable) or multidimensional (meaning that the questionnaire measures several relatively independent latent variables).

When determining dimensionality, the researcher may use an exploratory or confirmatory approach. In the first case, analyses performed on the collected data allow for identifying the number and nature of the dimensions measured by the tool without making prior assumptions. In the second case, the researcher posits a hypothesis regarding how many dimensions are measured by the instrument according to a theoretical model, and which items belong to which dimension. Then, statistical analyses are used to confirm or reject this hypothesis. This second approach is the focus of the present study.

In confirmatory analysis, confirmatory factor analysis (CFA) is applied to verify whether a theoretical model is well-fitted to the data. Typically, several alternative CFA models with varying numbers of dimensions (latent variables) are evaluated. To assess their fit, researchers use so-called fit indices, which assist in deciding which model best corresponds to the data.

One of the key issues in CFA is determining whether a given psychological construct is unidimensional or multidimensional - that is, whether it is driven by a single, coherent latent factor reflecting a homogeneous psychological trait, or not. Despite the widespread use of classical (global) fit indices, the literature does not provide definitive answers as to whether these coefficients can accurately differentiate between unidimensional and multidimensional structures, especially when the data were generated or collected for a clearly unidimensional construct. As a result, researchers may draw incorrect conclusions about the nature of the measured construct, which in turn affects result interpretation and the quality of diagnostic tools used.

The aim of this dissertation is to address this gap and answer three research questions: (P1) Do classical fit indices differentiate between unidimensional and multidimensional models when the data have a clearly unidimensional structure? (P2) Which fit indices are most useful

for assessing the accuracy of a CFA model? (P3) Is it possible to establish stable and practical cutoff points that would allow for valid conclusions about model dimensionality? To answer these questions, two studies were conducted: a simulation study (*N* = 31,500 datasets) and an empirical study (*N* = 5,172 participants). In both, theoretical models consistent and inconsistent with a unidimensional data structure were compared. The validity of the most commonly used indicators in the literature was analyzed: global fit indices (χ^2, χ^2/ df, χ^2_p, CFI, TLI, GFI, AGFI, RNI, RMSEA, AIC, BIC, RMR, and SRMR) as well as less common ones based on *modification indices* - local specification errors (*MI*, *NCP*, *ECP*, *m*, *nm*, *EPC:m*, *EPC:nm*, and *i*).

In the simulation study, data sets were generated with a clearly unidimensional structure, on which models correctly specified as unidimensional and incorrectly specified as multidimensional were compared. The goal was to replicate a situation in which the actual latent structure is unidimensional, but the researcher erroneously assumes that the construct has a more complex structure. The empirical analyses followed a similar approach, using data from real psychological studies employing theoretically unidimensional instruments. As in the simulated data, the aim was to assess whether classical fit indices and modification indices allow accurate differentiation between correct and incorrect models, considering typical levels of measurement error, noise, and individual variability found in empirical data.

The results obtained from both the simulation and empirical studies were consistent and showed that classical fit indices do not allow reliable differentiation between unidimensional and multidimensional models when the data are clearly unidimensional. (P1) Differences between the models were negligible and often random, which means that traditional fit indices are not effective in identifying the correct model. (P2) Only indices based on modification indices - particularly the nm% index - allowed for relatively clear assessments of model dimensionality. (P3) Attempts to determine stable and practical cutoff values for automatically classifying a model as unidimensional or multidimensional were unsuccessful.

The study points to the limited utility of classical fit indices in the analysis of unidimensional models and highlights the potential of an approach based on local specification error analysis. It proposes an alternative path involving the use of modification indices - unlike the global fit indices mentioned above, these are used to assess the correctness of specific parts of the model (such as individual paths or the assignment of questionnaire items to factors). The analyses conducted allowed for formulating recommendations regarding the use of modification indices as an alternative approach, which demonstrates substantial potential for

accurately distinguishing cases where only a single strong factor is present and doubts arise concerning the structure of the operationalized construct.

Keywords: confirmatory factor analysis, structure, simulations, fit coefficients, latent variable

1. WPROWADZENIE TEORETYCZNE

1.1. Pomiar cech latentnych w psychologii

W psychologii analizowane zmienne (konstrukty psychologiczne) często mają charakter latentny, czyli nieobserwowalny bezpośrednio; o takich cechach wnioskuje się na podstawie obserwowalnych danych, które mierzone są odpowiednio przygotowanymi narzędziami (Raykov i Marcoulides, 2011; Urbina, 2014). Narzędzia pomiarowe cech latentnych w dużej mierze sprowadzają się do metod kwestionariuszowych, najczęściej samoopisowych (Loewenthal i Lewis, 2020; Paulhus i Vazire, 2007). Takie narzędzia kwestionariuszowe składają się zwykle z serii stwierdzeń, względem których osoba badana określa swoje stanowisko (McDonald, 2013) za pomocą jakiejś skali odpowiedzi, na przykład od zdecydowanej zgodności do zdecydowanej niezgodności na skali Likerta (Loewenthal i Lewis, 2020; Robins, Fraley i Krueger, 2009). Te pytania nazywane są w języku polskim pozycjami kwestionariusza, pytaniami kwestionariuszowymi lub – za językiem angielskim – itemami (Hayes, 2021; Tanujaya, Prahmana i Mumu, 2022). Poszczególne itemy wyrażają różne aspekty cechy latentneji inólnie tworzą nadrzędne konstrukty lub wymiary psychologiczne odpowiadające danej właściwości psychologicznej (Kline, 2010; Rao i Sinharay, 2007). Przygotowanie takich kwestionariuszy wymaga spełnienia wymogów procedury psychometrycznej, aby narzędzie można było uznać za poprawnie mierzące to, co zakładają jego podstawy teoretyczne (Gulliksen, 2013; Kline, 2015). Ta procedura dotyczy kolejno definiowania obszaru mierzonej właściwości, następnie generowania pozycji testowych, a finalnie oceny, czy seria pytań kwestionariusza spełnia kryteria dobrego pomiaru (Cooper, 2023; Price, 2016; Urbina, 2014).

Określenie „dobry pomiar" zwykle dotyczy uzyskania satysfakcjonującej rzetelności i trafności narzędzia badawczego. Rzetelność narzędzia badawczego jest rozumiana zazwyczaj jako wielkość błędu pomiaru, oparta o powtarzalne i spójne wyniki przy wielokrotnym pomiarze, a trafność w najczęstszym ujęciu określa, w jakim stopniu narzędzie mierzy to, co zakładają jego podstawy teoretyczne (Kline, 2015, McDonald, 2013). Trafność i rzetelność w obrębie pojedynczego narzędzia badawczego mogą być ze sobą związane: wyższa trafność narzędzia badawczego może prowadzić do wyższego poziomu związków pomiędzy pytaniami kwestionariusza (wówczas w większym stopniu mierzą to samo), a ta z kolei do wyższego poziomu rzetelności (której miara w najprostszym ujęciu opiera się na średnim poziomie korelacji między pytaniami; zob. Rak i Wrześniowski, 2024). W nieco innym rozumieniu

pojęcia trafności mówi się o tzw. *trafności strukturalnej*, czyli zgodności struktury teoretycznej narzędzia psychometrycznego z rzeczywistą strukturą uzyskaną w danych empirycznych.

Jednym z problemów w ramach procedury tworzenia narzędzia badawczego jest rozstrzygnięcie wymiarowości konstruktu. Badany konstrukt może mieć teoretycznie strukturę jedno- lub wielowymiarową (Furr, 2021; Lee, 2001). Weryfikacja założeń teoretycznych co do liczby wymiarów danego zjawiska odbywa się na etapie, w którym zebrane zostaną dane empiryczne, to znaczy uzyskane zostaną odpowiedzi osób badanych (Barendse, Oort i Timmerman, 2015). Wówczas zbiór uzyskanych danych poddawany jest analizie statystycznej (El-Den, Schneider, Mirzaei i Carter, 2020; Lee, 2001).

Zasadniczo w zakresie struktury mierzonego konstruktu podczas budowania narzędzi pomiarowych stosuje się podejście empiryczne lub podejście teoretyczne (Byrne, 2005; Kyriazos i Poga-Kyriazou, 2023; Thompson, 2004; Yang, 2005). W podejściu empirycznym badacz zakłada, że nie zna z góry struktury badanego zjawiska, dlatego w toku analizy eksploracyjnej, zazwyczaj eksploracyjnej analizy czynnikowej (EFA, *exploratory factor analysis*) lub analizy głównych składowych (PCA, *principal component analysis*) odkrywa, ile i jakich wymiarów ma dana właściwość (Cudeck i MacCallum, 2007; Norris i Lecavalier, 2010; Raykov i Marcoulides, 2011). W podejściu teoretycznym badacz zakłada określoną strukturę mierzonej cechy, to znaczy określa z góry, ile wymiarów ma badany konstrukt i które pytania kwestionariusza wchodzą w skład którego wymiaru (Kline, 2023; Maydeu-Olivares i McArdle, 2005). To drugie podejście wykonuje się z użyciem modelowania strukturalnego, a konkretnie analizy konfirmacyjnej (*confirmatory factor analisys*, CFA; Kline, 2015; Konarski, 2009). Niniejsza praca skupia się na tym drugim podejściu.

1.2. Konfirmacja modelu teoretycznego cechy latentnej

Podejście konfirmacyjne (jak większość współczesnych analiz) realizowane jest w którymś z popularnych pakietów do analizy CFA (Baldwin, 2019; Collier, 2020; Khademi, 2017). Badacz ustala, jak jego zdaniem powinna wyglądać struktura, a więc ile jest wymiarów danego konstruktu, które pytania kwestionariusza łączą się w te wymiary oraz w jakich relacjach są ze sobą te wymiary (Civelek, 2018). Precyzyjnie rzecz ujmując ustalane są w obrębie całej analizowanej struktury tzw. *model ścieżek* (*path model*) i *model pomiarowy* (*measurement model*) (Rosseel, 2012). W *modelu pomiarowym* badacz określa, w jaki sposób obserwowalne wskaźniki (tu: pytania kwestionariusza) odzwierciedlają konstrukty latentne (np. wymiar samooceny); z kolei w *modelu ścieżek* określa relacje między zmiennymi latentnymi

(tu: wymiarami jednego kwestionariusza)[1]; *ścieżkami* zaś nazywane są informacje o tym, jak poszczególne zmienne wchodzą ze sobą w relacje (tu: czy dwie zmienne w ogóle są jakoś związane, a jeśli tak, to czy tylko ze sobą korelują czy też jedna jest predyktorem drugiej, lub składową wymiaru latentnego; Konarski, 2009). W ujęciu statystycznym chodzi o tzw. specyfikację szczegółową równania SEM (*structural equation modeling*), które sprowadza się do zapisu matematycznego równania opisującego zmiennymi i ich współczynnikami regresji wszystkie zależności *modelu ścieżek* i *modelu pomiarowego* (Blunch, 2012). Mimo złożoności całego procesu, dzięki rozwojowi oprogramowania użytkowego dla psychologów, współcześnie w pakietach statystycznych ta specyfikacja jest wykonywana z użyciem wizualnych interfejsów z częściową automatyzacją niektórych kroków specyfikacji, często jednocześnie łącząc różne elementy *modelu pomiarowego* i *modelu ścieżek* w jedną wizualną reprezentację (Blunch, 2012; Finchi in., 2016).

Po wyspecyfikowaniu modelu badacz oblicza z użyciem oprogramowania tak zwane miary dopasowania modelu, które określają czy model jest poprawny (dobrze dopasowany) czy niepoprawny (źle dopasowany) względem danych empirycznych zebranych od osób badanych (Kline, 2023; Konarski, 2009; Westland, 2015). Innymi słowy, badacz wspomagając się oprogramowaniem statystycznym generuje serię wyników obliczeń wynikających (bezpośrednio lub pośrednio) z równań SEM, a następnie na podstawie różnych zaleceń wykazuje, czy jego model teoretyczny w świetle uzyskanych wyników okazał się poprawny (Bowen i Guo, 2011; Collier, 2020; Mair, 2018).

Oprócz sytuacji, w której badacz ma tylko jeden model teoretyczny i w obrębie zebranych danych chce sprawdzić, czy taki model jest poprawny (dopasowany) czy nie, pojawiają się sytuacje gdzie dla tych samych danych zachodzi potrzeba przetestowania kilku różnych modeli (Kline, 2023). Badacz może nie być pewnym, ile dokładnie jego konstrukt ma wymiarów, bo jego teoria zakłada alternatywne wyjaśnienia zjawiska. Badacz może popełniać błędy teoretyczne prowadzące do złych założeń względem modelu, albo po prostu chcieć wykazać, że jego model jest najlepszym z możliwych modeli teoretycznych (Lin, Huang i Weng, 2017; Preacher i Merkle, 2012). W każdym z tych przypadków głównym zadaniem w podejściu teoretycznym jest wybór właściwego modelu (West, Taylor i Wu, 2012). W każdej z wymienionych sytuacji w praktyce po prostu oblicza się współczynniki dopasowania kilku modeli, a następnie je porównuje (zob. Barendse i in., 2015; Schmitt, Sass, Chappelle

[1] Naturalnie, w przypadku jeszcze bardziej złożonych modeli łączących analizy CFA i ogólne modelowanie SEM mogą pojawić się w modelu ścieżki między zmiennymi latentnymi a zmiennymi obserwowalnymi (na przykład między samooceną jako zmienną latentną tworzoną z pytań kwestionariusza a obserwowalnym bezpośrednio wzrostem osoby badanej).

i Thompson, 2018). Można wówczas określić, który model teoretyczny jest lepszy, a więc bardziej trafny w świetle zebranych danych.

1.3. Nieprawidłowa specyfikacja modelu i jej konsekwencje

W przypadku, gdy struktura danego zjawiska z poziomu teorii jest trudna do określenia, badacz w praktyce może testować różne alternatywne modele (Wang i Wang, 2019). Poszczególne teoretyczne modele alternatywne mogą się różnić w nieznacznym stopniu, na przykład kwestią tego, ile pytań wchodzi w obręb danego wymiaru, ale mogą też różnić się fundamentalnie, np. ile właściwie wymiarów ma badane zjawisko, czy istnieje wymiar nadrzędny itd. (André, Munck, Håkansson i Claesdotter-Knutsson, 2022; Byrne, 2005; Shevlin i Adamson, 2005; Yu, Lin i Hsu, 2013). Można wprawdzie uznać, że lepszy jest ten model, który jest lepiej uzasadniony teoretycznie, ale bywa przecież również tak, że uzasadnienie teoretyczne może zawierać błędy.

Możliwość popełniania przez badacza błędów w obrębie założeń modelu jest ogólnie określana mianem nieprawidłowej specyfikacji (*model misspecification*; Evermann i Tate, 2010; Rhemtulla, 2016; Robitzsch, 2023). Jest to obecnie dość szerokie pojęcie, mogące dotyczyć wielu różnych błędów, które literatura dzieli na tzw. błędy wewnętrzne i błędy zewnętrzne. Wewnętrzne błędy modelu to błędy polegające na wyborze (lub braku) istotnych dla modelu pomiarowego układu (wzajemnych relacji) zmiennych (tu: pytań kwestionariusza), czy błędnym rozwiązaniu czynnikowym w modelu (Bollen i Noble, 2011; Dennis, Ponciano, Taper i Lele, 2019; Pereira i Cribari-Neto, 2014). Z kolei zewnętrzne błędy modelu to błędnie stosowane algorytmy (np. wybrana nieadekwatna metoda obliczenia współczynników), problemy mocy statystycznej modeli, możliwe przeszacowanie lub niedoszacowanie parametrów (na przykład nieprawidłowe szacowanie parametrów modelu z uwagi na błędną jego specyfikację matematyczną; Hayashi, Bentler i Yuan, 2007; Iacobucci, 2010; Rhemtulla, 2016; Wang i Rhemtulla, 2021; Yuan i Bentler, 2007). W efekcie każdego z tych problemów zadany model teoretyczny nie będzie dobrze korespondował z danymi empirycznymi, zatem odpowiednie współczynniki powinny wskazać, że model jest po prostu błędny. W literaturze poddaje się jednak w wątpliwość kwestię, na ile modele SEM są realnie odporne na różne typy nieprawidłowej specyfikacji (por: Evermann i Tate, 2010; Fiski in.,2023; Henseler, 2012; Robitzsch, 2023), ale zwykle dotyczy to po prostu krytyki samej podstawy modelowania SEM, a nie rozstrzygania o poprawności modeli poprzez statystyki dla nich uzyskane. Ponadto, debata ta jest prowadzona zasadniczo już poza psychometrią – w obrębie artykułów w

czasopismach statystycznych i matematycznych – i w dużej mierze, z punktu widzenia badań psychologicznych, jest po prostu niepraktyczna z uwagi na brak rozstrzygających dowodów matematycznych, a co za tym idzie nie ma wyraźnego powodu by psychometria jako taka rezygnowała z obecnie stosowanych narzędzi statystycznych (zob. Borsboom, 2006).

Co ważne, we współczesnym ujęciu nieprawidłowej specyfikacji zakłada się, że każdy model statystyczny jest do pewnego stopnia błędny, stąd dopuszcza się, a nawet zaleca, nieidealne rozwiązania, jeśli są one zgodne z teorią (na przykład: Dennis i in., 2019; Robitzsch, 2023; Hayashi, Bentler i Yuan, 2007). I choć możliwe powody nieprawidłowej specyfikacji mogą być różne, ochroną przed błędnym wyborem modelu ma być odpowiednie użycie współczynników dopasowania modelu, które wykażą, że model po prostu jest błędny (Cerreia-Vioglio, Hansen, Maccheroni i Marinacci, 2020; Raykov i Marcoulides, 2011). Znów więc należy zwrócić uwagę, że kiedy realna struktura zjawiska nie jest znana, rozstrzygnięcie, który model jest najlepszy, jest wypadkową pomiędzy tym, na ile badacz poprawnie sformułował teorię, a tym, na ile uzyskane współczynniki statystyczne są satysfakcjonujące względem konkretnego modelu, w związku z czym „najlepszy model" nie musi oznaczać, że jest on poprawny.

W świetle współczesnej literatury przedmiotu uprawnionym wydaje się założenie, że wnioski dotyczące trafności modelu teoretycznego (poprawności względem danych empirycznych) powinny być jednoznacznie rozstrzygane na podstawie wyników analiz statystycznych, a tam gdzie model teoretyczny po prostu jest niezgodny z danymi współczynniki pozwolą na jego odrzucenie. Ponieważ jednak między modelami traktowanymi jako poprawne i niepoprawne w publikacjach można znaleźć jedynie minimalne różnice na przykład na trzecim miejscu po przecinku (zob. także Preacher i Merkle, 2012; Schmitt, Sass, Chappelle i Thompson, 2018), pojawia się także uzasadniona wątpliwość, czy w takich wypadkach badacz podejmuje decyzje w sposób jednoznaczny, oraz czy badacz może błędnie przyjmować dany wynik za pewną i poprawną podstawę wnioskowania. Niniejsza praca dotyczy więc tego, czy i w jakich wypadkach badacz, wybierający ścieżkę teoretyczną (konfirmacyjną) i polegając na dostępnych w literaturze wskaźnikach dopasowania, może podjąć w zakresie oceny poprawności modelu jednoznaczną i pewną decyzję.

Rozważania zawarte w pracy dotyczą zatem kwestii oceny (z użyciem stosowanych w psychologii metod konfirmacyjnych), czy liczba wymiarów struktury operacjonalizowanego zjawiska jest poprawna. Na przykład, nawet jeśli w istocie realna struktura konstruktu jest jednowymiarowa (Timmerman, Lorenzo-Seva i Ceulemans, 2018), badacz nadal, na poziomie budowania teorii lub narzędzia, może uznać, że dany konstrukt ma więcej niż jeden wymiar.

Jak obecnie rozstrzyga się, czy taka hipoteza jest potwierdzona? W przypadku struktur jednowymiarowych wynikających z teorii badacze przyjmują, że taka struktura powinna mieć wyraźnie lepszy poziom dopasowania w analizie CFA niż złożona z tych samych pytań struktura wielowymiarowa (czyli dwuwymiarowa, trójwymiarowa i tak dalej; np.: Rytilä-Manninen i in., 2016; Strelhow, Sarriera i Casas, 2020). Niestety, obecna literatura przedmiotu nie daje jednoznacznych odpowiedzi, czy taka metoda jest poprawna. Konkretniej, **nie zostało dotychczas empirycznie stwierdzone, czy rzeczywiście współczynniki dopasowania różnią się istotnie pomiędzy modelami jedno- i wielowymiarowymi utworzonymi na tych samych danych o niewątpliwie jednowymiarowej strukturze. Będzie to zatem pierwszy główny obszar badań niniejszej pracy doktorskiej i pierwsze pytanie badawcze (P1).**

Dlaczego nieprawidłowa specyfikacja modelu jest problemem? Przede wszystkim, kiedy badacz tworzy narzędzie kwestionariuszowe do badania cechy latentnej (konstruktu psychologicznego), uznanie modelu błędnego za poprawny prowadzi do równie błędnego wyobrażenia o strukturze cechy, którą badacz chce operacjonalizować (np.: Henley, Golden i Kashner, 2020; Jarvis, MacKenzie i Podsakoff, 2012). Pojawia się także pewien problem natury statystycznej, gdy za prawidłowy zostanie uznany model, który jest redundantny (ma zbyt dużo wymiarów). W takiej sytuacji zazwyczaj mamy do czynienia z silnymi korelacjami pomiędzy jego wymiarami – są one redukowalne i zawierają wspólną wariancję (Sun, Konold i Fan, 2011; zob. także: Dalal, 2023; Headey, 2008). Z kolei duże korelacje (tzw. nadmierna współliniowość) pomiędzy wymiarami tego samego kwestionariusza prowadzą między innymi do nieprawidłowości w modelach statystycznych opartych o analizę regresji (Alin, 2010; Ayanso i Mokaya 2013; Sifuentes-Amaya i Ramirez-Valverde, 2010), a konkretniej do błędów w ocenie związków i wpływów (Daoud, 2017; Kim, 2019; Winship i Western, 2016). Jak wykazują różni badacze, pojawienie się nadmiernej współliniowości może skutkować tym, że analizy regresji źle szacują wartości współczynników regresji – tam, gdzie realnie istnieje istotna zależność, model może zwracać informację o jej braku lub odwracać znaki współczynników, a w efekcie prowadzić do błędnej interpretacji wyników (Kalnins, 2018; Kalnins, 2022; Komlos, 2019; Tay, 2017; Singh, Singh i Paprzycki, 2023). Innymi słowy, badacz wykorzystując wadliwy model o nadmiernej liczbie wymiarów (obciążony błędem redundancji) może w konsekwencji nie dostrzegać jakichś zależności lub mylnie je interpretować (np. odwrotnie niż jest w rzeczywistości). Dlatego ważne jest określenie, z jak dużą pewnością badacze mogą szacować wymiarowość badanych konstruktów, kiedy używają klasycznych metod oceny modelu.

1.4. Metody (miary) oceny poprawności modelu teoretycznego w analizie SEM/CFA

W ocenie poprawności modelu wykorzystuje się powszechnie pewne współczynniki ogólnie nazywane *miarami dopasowania*, jednak obecnie nie ma konsensusu na temat tego, które z nich są najlepsze (Konarski, 2009; Putnick i Bornstein, 2016), ani nawet czy któraś z proponowanych reguł akceptacji modelu poprzez wartości tych współczynników jest lepsza od innych (Markland, 2007; Niemand i Mai, 2018; Sivo, Fan, Witta i Willse, 2006; Wolf i McNeish, 2023). Zarazem jednak w praktyce badawczej i publikacyjnej niektóre z nich są zdecydowanie częściej raportowane (Greiff i Heene, 2017; Iacobucci, 2010; Jackson, Gillaspy Jr. i Purc-Stephenson, 2009; Kenny, 2015). Miary te zostaną poniżej jedynie pokrótce opisane, ponieważ przedmiotem zainteresowania niniejszej pracy nie jest ich szczegółowa konstrukcja matematyczna, ale rzeczywiste funkcjonowanie na danych empirycznych. Bardziej szczegółowy opis znajduje się w pracach na temat modelowania strukturalnego (np. Kline, 2015; Konarski, 2009; Marcoulides i Schumacker, 2013; Sideridis i Jaffari, 2022) oraz w publikacjach opisujących rekomendacje odnośnie ich interpretacji (rekomendacje bardziej liberalne: Sharma i in., 2005; rekomendacje bardziej konserwatywne: Sivo i in., 2006). Najogólniej rzecz biorąc, miary te można podzielić na grupy wymienione poniżej

1.4.1. Miary globalne

Pierwszą grupę miar stanowią globalne miary dopasowania (Hooper i in. 2008; Jackson, Gillaspy i Purc-Stephenson, 2009). Mamy tu: (a) miarę χ^2 wraz z raportowaną czasem miarą χ^2/df i istotnością statystyczną miary χ^2 (czyli χ^2_p), (b) miary RMSEA, GFI, AGFI, RMR i SRMR, które określane są jako miary dopasowania absolutnego (*absolute fit indices*) – to znaczy miary, które określają stopień dopasowania całego modelu do danych bez porównywania go do innych modeli, oraz (c) TLI, CFI i RNI, które są miarami dopasowania przyrostowego (*incremental fit indices*). Zostaną teraz kolejno omówione wszystkie wspomniane miary globalne.

Miara χ^2 ocenia, na ile model teoretyczny (wybrany do testowania przez badacza) jest zgodny z obserwowanymi danymi. Na poziomie algorytmu porównuje ona macierz kowariancji uzyskaną z modelu teoretycznego z macierzą kowariancji obliczoną na podstawie rzeczywistych danych. Ponieważ wyższe wartości wskazują na różnice między modelem teoretycznym a rzeczywistymi danymi, pożądana jest jak najniższa wartość tej miary (Zheng i Bentler, 2024), przy czym nie ma rekomendowanej jej górnej czy dolnej granicy. Do względnej oceny tego współczynnika może posłużyć poziom jego istotności statystycznej χ^2_p,

który pozwala ocenić, czy różnice między macierzą kowariancji uzyskaną z modelu a macierzą kowariancji z realnych danych są istotne statystycznie (pożądany jest poziom powyżej wartości 0,05). Jednak – jak wskazują liczne artykuły (np. Bollen, 1990; Kyriazos, 2018) – wraz ze wzrostem wielkości próby ta statystyka prawie zawsze jest istotna, a więc jej wartość diagnostyczna w dużych próbach może być niewielka. Pośrednim rozwiązaniem jest raportowanie wartości χ^2/df czyli stosunku surowej wartości χ^2 do liczby stopni swobody w modelu (*df, degrees of freedom*, tzn. liczba niezależnych informacji dostępnych do oszacowania parametrów modelu, które rozumie się jako nadmiar danych w stosunku do liczby oszacowywanych parametrów)[2], która ma na celu skorygować wielkość wartości o liczbę zmiennych w modelu (Kline, 2023). W literaturze wprawdzie są pewne rekomendacje dotyczące progu akceptowalności modelu, ale są one niejednoznaczne, ponieważ w części publikacji za akceptowalne wskazuje się miary poniżej 5 (Hu i Bentler, 1999), w niektórych poniżej 3 (Schumacker i Lomax, 2016), a w niektórych poniżej 2 (Kline, 2015), przy czym z jednej strony szuka się wartości przeciętnych, dla których χ^2 po prostu nie uzyskiwałby istotności statystycznej, a z drugiej - szuka się wskazówek co do tego, który poziom χ^2/df po prostu mógłby wskazywać na model warty akceptacji (Civelek, 2018). Jako *akceptowalny* próg dla tej miary w obrębie niniejszej pracy zostanie przyjęta wartość poniżej 5, zgodnie z rekomendacjami Hu i Bentler (1999).

Miarą matematycznie opartą o surową wartość χ^2 jest także RMSEA (*root mean square error of approximation*, średni błąd kwadratowy aproksymacji), która mierzy błąd aproksymacji z uwzględnieniem stopni swobody, co oznacza, że bierze pod uwagę złożoność modelu oceniając jego dopasowanie (Konarski, 2009). Choć w najczęstszym zaleceniem dla dobrych modeli jest wartość tego wskaźnika poniżej 0,05 (np. Steiger, 2007), to w publikacjach często dopuszcza się jako akceptowalne modele dla których wartość jest mniejsza niż 0,10 (np. Bedyńska, 2012).

Kolejne wskaźniki to indeks dobroci dopasowania GFI (*goodness-of-fit index*) oraz skorygowany indeks dobroci dopasowania AGFI (*adjusted goodness-of-fit index*) autorstwa Jöreskoga i Sörboma (1993). Są to wskaźniki dopasowania bezwzględnego, które informują, jak dobrze model pasuje do danych (porównując obserwowane macierze kowariancji z tymi przewidywanymi przez model), przy czym AGFI jest skorygowany o liczbę parametrów, co czyni go bardziej rygorystycznym dla bardziej złożonych modeli (Konarski, 2009). Wskaźniki

[2] W praktyce modelowania SEM jest to po prostu wartość *df* = liczba niezależnych obserwacji – liczba szacowanych parametrów, czyli: liczba osób badanych – liczba parametrów do oszacowania (elementów modelu takich jak współczynników ścieżkowych, kowariancji, błędów pomiaru itd.)

te w formie ułamków dziesiętnych, zawarte w przedziale od 0 do 1, powinny dążyć do wartości jak największej (najbliższej 1), przy czym współcześnie rekomenduje się przekroczenie wartości 0,950, choć w starszych publikacjach można znaleźć propozycje akceptacji także wartości powyżej 0,800 (Sharma i in., 2005; Sivo i in., 2006).

Podczas gdy GFI i AGFI są bardziej odpowiednie do oceny ogólnego dopasowania modelu w kontekście liczby zmiennych (większy nacisk jest położony w równaniu na korektę o liczbę zmiennych), dwie kolejne miary, RMR (*root mean square residual*) i SRMR (*standardized root mean square residual*, standaryzowana miara RMR) koncentrują się na bezpośrednich różnicach między danymi a modelowanymi macierzami kowariancji (ściślej, obliczenia dotyczą standaryzowanej różnicy między zaobserwowanymi a przewidywanymi związkami; Maydeu-Olivares, 2017). Im mniejsze wartości tych dwóch współczynników tym lepsze dopasowanie modelu, a wartość zerowa oznacza idealne dopasowanie, przy czym wartość mniejsza niż 0,08 dla SRMR jest ogólnie uważana za dobre dopasowanie (Hu i Bentler, 1999), a dla RMR przyjmuje się zwykle za Sivo i in. (2006) wartość poniżej 0,02.

Ostatnie wskaźniki w tej grupie to względny indeks dopasowania CFI (*comparative fit index*), współczynnik Tucker-Lewisa TLI (*Tucker-Lewis index*) oraz RNI (*relative noncentrality index*)[3]. Stanowią one uzupełnienie informacji o modelu. Wszystkie trzy wskaźniki są wskaźnikami przyrostowymi (*incremental fit indices*), a więc w przeciwieństwie do poprzednio wymienionych dostarczają informacji o tym, jak dobrze cały model jest dopasowany do danych względem jednego z trzech możliwych punktów odniesienia: a) modelu zerowego (*null model*), który zakłada całkowity brak zależności między zmiennymi obserwowalnymi, gdzie wszystkie kowariancje są równe 0; b) modelu niezależnego (*independence model*), który również zakłada brak zależności między zmiennymi, ale zwykle uwzględnia wariancje każdej zmiennej, lub c) modelu bazowego (*baseline model*) który w zależności od oprogramowania[4] może być tożsamym z modelem zerowym lub niezależnym, ale w niektórych podejściach może być modelem minimalnym, czyli na przykład modelem jednoczynnikowym. Wszystkie trzy uwzględniają też złożoność modelu (CFI w największym stopniu), a więc faworyzują modele o niższej złożoności (Konarski, 2009). Taki zabieg

[3] Niecentryczność (lub w niektórych tłumaczeniach niecentralność) uwzględniania w obrębie metody RNI może być także rozumiana jako stopień odchylenia modelu od idealnego dopasowania, czyli sytuacji, w której wszystkie zmienne są całkowicie wyjaśnione przez model zaproponowany przez badacza.

[4] Na przykład w dokumentacji funkcji *fitMeasures* pakietu *lavaan* wykorzystywanego w niniejszej pracy znajduje się parametr *baseline.model*, który pozwala użytkownikowi zdefiniować własny model bazowy (czyli użytkownik definiuje jak rozumie model bazowy). Jeśli jednak użytkownik nie dostarczy własnego modelu bazowego, *lavaan* automatycznie wybiera tu model niezależny i wykorzystuje go jako punkt odniesienia przy obliczaniu takich miar dopasowania jak CFI czy TLI. Podobną informację można znaleźć w dokumentacji Mplus. IBM AMOS traktuje jako tożsame model bazowy i model zerowy.

w praktyce pomaga uniknąć sytuacji, w której bardziej złożone modele są preferowane tylko dlatego, że lepiej pasują do danych pochodzących od konkretnej próby, ale niekoniecznie są najlepszymi modelami teoretycznymi. Zatem idea stojąca za ich interpretacją określa nie tyle, na ile model jest dopasowany do danych, ale na ile lepszym modelem jest ten zaproponowany przez badacza (względem modelu bez związków między zmiennymi), zatem można rozumieć je jako w większym stopniu potwierdzenie założonych związków niż dopasowanie do struktury samych danych. Na poziomie metod ich obliczania wszystkie trzy w uproszczeniu dokonują oceny modelu bazowego i modelu teoretycznego poprzez uzyskanie dla nich wartości χ^2, a następnie – z uwzględnieniem stopni swobody obu modeli – oceniają poziom różnicy między nimi sprowadzając ją do wartości ułamkowej. Wartości w przedziale od 0 do 1 mówią o poziomie dopasowania, przy czym im wyższe, tym lepsze jest dopasowanie, a w publikacjach dla wszystkich trzech znajdziemy zarówno liberalne progi akceptacji: powyżej 0,80 (Sharma i in., 2005) jak i konserwatywne, sugerujące wartości powyżej 0,95 (Taasoobshirazi i Wang, 2016; Ximénez i in., 2022).

Miary globalne mogą wskazać, że model jest błędny, ale zwykle nie wskazują bezpośredniego powodu błędu (Konarski, 2009). Problemem miar globalnych jest to, że w praktyce zaleca się jednoczesną interpretację kilku z nich, co w niejednoznacznych sytuacjach (na przykład jeden wskaźnik sugeruje dobre dopasowanie, inny złe) może znacznie utrudniać weryfikację modelu (Fan i Sivo, 2005).

1.4.2. Miary przyrostowe

Druga grupa miar wykorzystanych z pracy to popularne miary różnic między modelami. Dla dwóch modeli strukturalnych można dokonać porównania ich surowych wartości χ^2, czyli obliczyć miarę różnicy między ich wartościami χ^2_{diff} wraz z istotnością statystyczną tej różnicy $\chi^2_{diff\text{-}p}$, która ma określić, czy modele są rozróżnialne, a w konsekwencji - czy któryś może być lepszy (Werner i Schermelleh-Engel, 2010). W tej grupie miar porównawczych znajdują się też wskaźniki AIC (*Akaike information criterion*) oraz BIC (*bayesian information criterion*) (Vrieze, 2012). Są to możliwe do obliczenia względne parametry jakości dopasowania modelu do danych oparte o algorytm maksymalnej wiarygodności[5]. Ponieważ dla każdego modelu istnieje właściwie tylko jedno (najlepsze)

[5] Maximum Likelihood Estimate (MLE), czyli estymacja maksymalnej wiarygodności to metoda statystyczna która (we spółczesnym oprogramowaniu) automatyzuje poszukiwanie najlepszych parametrów modelu, w taki sposób, aby zwiększyć funkcję wiarygodności, czyli prawdopodobieństwo uzyskania obserwowanych danych przy założonym określonym modelu. W kontekście SEM, ta funkcja po prostu opisuje, jak prawdopodobne są

rozwiązanie algorytmu AIC i BIC, można przyjąć, że różnice między nimi dla alternatywnych modeli pozwalają badaczom wybrać lepszy model, czyli ten, który ma lepsze parametry jednego z tych wskaźników lub obu (Chakrabarti i Ghosh, 2011; Lin, Huang i Weng, 2017). Problemem tych dwóch wskaźników jest to, że w przypadku prób o dużej liczebności, BIC może preferować prostsze modele, podczas gdy AIC może wskazywać na złożone modele jako lepsze. Jeśli oba kryteria prowadzą do tego samego wniosku (wskazują ten sam model), to można mieć większą pewność, że wybrany model jest rzeczywiście najlepszy. W innych przypadkach badacz może zadecydować pomiędzy wyborem jednego z nich: konserwatywnego BIC lub elastycznego AIC (Kuha, 2004). Obie miary nie mają górnego limitu, są po prostu wartościami większymi od 0.

1.4.3. Indeksy modyfikacji

Ostatnia grupa miar dopasowania `to alternatywne podejście, które nie jest oparte o pojedyncze współczynniki dla całego modelu. Saris, Satorra i Van der Veld (2009) zwrócili uwagę na kwestię złożonej interpretacji jakości modelu strukturalnego poprzez globalne współczynniki dopasowania. Ich zdaniem takie współczynniki dopasowania mogą wskazywać, że model jest nieprawidłowy, ale nie pozwalają na określenie, który z nieprawidłowych modeli jest obarczony najmniejszym błędem. W związku z tym zaproponowali podejście oparte o tak zwane indeksy modyfikacji (*modification indices, MIs*) pozwalające określić lokalne błędy w specyfikacji (*local misspecifications*).

Indeksy modyfikacji to informacje dotyczące każdego z *obiektów* modelu konfirmacyjnego, na przykład osobnych ścieżek tworzących z poszczególnych pytań kwestionariusza zmienną latentną (Kim, 2011; Whittaker, 2012). W przeciwieństwie do omówionych wcześniej globalnych miar nie dotyczą więc one całego modelu jako takiego, ale odnoszą się do (na przykład) poszczególnych pytań ujętych w kwestionariuszu, a konkretnie tego, w jakim stopniu dane pytanie jest poprawnym wskaźnikiem jakiejś zmiennej latentnej. Dla każdego takiego obiektu obliczane są pewne wskaźniki, które pozwalają ocenić, czy dodanie lub usunięcie którejś ze ścieżek mogłoby poprawić jakość modelu (Sarris i in., 2009). Współcześnie wykorzystywane oprogramowanie statystyczne dostarcza badaczowi automatycznie wyników tych obliczeń, testując od razu większość możliwych modyfikacji ścieżek zarówno istniejących, jak i potencjalnych, czyli nieprzewidzianych przez testowany

---

obserwowane dane przy założeniu, że model jest prawdziwy. Funkcja wiarygodności jest obliczana na podstawie różnic między macierzą kowariancji obserwowanej a przewidywaną przez model.

model (Clayton i Pett, 2008; Rosseel, 2012; Shek i Yu, 2014). Badacz testując model, przeglądając wartości indeksów modyfikacji dla ścieżek istniejących w modelu, może sprawdzić, czy daną ścieżkę należy usunąć, ponieważ obniża parametry modelu, a dla ścieżek potencjalnych, czy należy dodać daną ścieżkę, której w modelu nie uwzględniano, a bez niej model jest gorzej dopasowany. Podejście to, zawierające pewne elementy eksploracyjne, w obrębie metod SEM/CFA bywa krytykowane, ponieważ może tworzyć modele oparte na danych, a nie teorii, a zatem nie znajdujące potem potwierdzenia w badaniach na innych próbach (zob. Hoyle, 2012). W niniejszej pracy zostaną jednak wykorzystane również te wskaźniki, ponieważ ich odpowiednie zastosowanie pozwala wykorzystać ich zalety i uniknąć problemów.

Z surową wartością *MI* powiązane są współczynniki *NCP* i *ECP*. Funkcja tych wszystkich trzech podstawowych współczynników modyfikacji jest ta sama – mają za zadanie wskazać czy dodanie lub usunięcie którejś ze ścieżek mogłoby poprawić jakość modelu, niemniej wyrażane są poprzez nieco inne metody matematyczne (Sarris i in., 2009). *MI* to klasyczna już miara działająca w oparciu wartość χ^2 modelu. Z matematycznego punktu widzenia sprowadza się do analizy, czy na przykład usunięcie lub dodanie jakiejś ścieżki (obiektu w modelu) wpłynie na poprawę (lub pogorszenie) dopasowania modelu do danych (Fuoli, 2022; Kim, 2011) poprzez sprawdzenie różnic między obecną (wynikającą z testowanego modelu) wartością surową χ^2, a wartością surową χ^2 dla modelu z uwzględnioną zmianą.

Współczynnik *NCP* (*noncentrality parameter,* parametr niecentralności) wskazuje, jak dobrze model (potencjalny) pasuje do danych pod względem odchyleń od założonej struktury (Hancock i Mueller, 2013). Konkretniej, dla pojedynczej ścieżki wyższa wartość *NCP* wskazuje na znaczne różnice między przewidywaniami modelu a rzeczywistymi danymi w obrębie tej ścieżki, czyli jest to (w uproszczeniu) ocena dopasowania pojedynczej ścieżki do danych (w tym ścieżki potencjalnej, której jeszcze w modelu nie ma).

Z kolei *EPC* (*expected parameter change,* oczekiwana zmiana parametru) informuje w obrębie modelu o przewidywanej przeciętnej zmianie ładunków czynnikowych dla pozostałych ścieżek (Hancock i Mueller, 2013), a zatem zamiast odnosić się do stopnia dopasowania modelu odnosi się do zmiany wielkości współczynników w modelu.

Wszystkie trzy współczynniki (*MI*, *NCP*, *ECP*) nie mają wartości maksymalnej, a ich oceny i porównywania zwykle dokonuje się tylko w obrębie pojedynczego testowanego modelu, tzn. badacz w obrębie obliczonych współczynników wybiera najlepszą potencjalną ścieżkę. W praktyce oprogramowanie (np. IBM SPSS AMOS, R, laavan) oblicza wartości tych

trzech parametrów dla wszystkich istniejących i nieistniejących ścieżek, badacz modyfikuje model i ponownie oblicza wszystkie współczynniki, jeśli chce wykonać kolejny krok. Wysokie wartości współczynników MI i ECP sugerują, że zmiana którejś ścieżki (jej dodanie lub usunięcie) mogłoby znacząco poprawić dopasowanie modelu (jego zgodność z danymi empirycznymi), a niska wartość NCP sugeruje lepsze dopasowanie modelu po wprowadzeniu danej zmiany.

Dla indeksów modyfikacji w tym ujęciu nie ma jednoznacznych wartości referencyjnych (*cutoff* lub *rule of tumb*), jest raczej to zbiór zaleceń w postaci „jeśli dane parametry mają wartość X i Y, to wartość referencyjna wynosi Z, a to oznacza konkluzję K", stąd ich wymienianie w formie zbliżonej do formy wartości referencyjnej klasycznych miar dopasowania nie wydaje się być szczególnie prawidłowe. W związku z tym Saris, Satorra i Van der Veld (2009) zaproponowali algorytm interpretacji wartości *MI*, *EPC* oraz *NCP*, czyli zbiór reguł dla tych trzech parametrów z wartościami odcięcia dla poszczególnych decyzji Algorytm ten został oprogramowany w języku R przez Uanhoro (2017).

Zgodnie z tym algorytmem jest pięć możliwości: 1) ścieżka jest niepoprawnie wyspecyfikowana (*misspecified*) w oparciu o wartości *MI* i *NCP*, (zapisywane jako *m*), 2) dla ścieżki nie ma nieprawidłowej specyfikacji (*nm*, czyli nie można wskazać, że jest niepoprawnie wyspecyfikowana), 3) ścieżka jest niepoprawnie wyspecyfikowana (*misspecified*), ale wynika to tylko z wartości EPC (*EPC:m*); 4) dla ścieżki nie ma nieprawidłowej specyfikacji z użyciem wartości EPC (*EPC:nm*) lub 5) analiza jest niekonkluzywna dla danej ścieżki (*i*).

Biorąc pod uwagę interpretację konkluzji dla ścieżek należy zauważyć, że dotyczą one zarówno ścieżek istniejących, jak i potencjalnych. Oznacza to, że jeśli model jest poprawnie wyspecyfikowany (jego struktura teoretyczna jest bliższa strukturze danych empirycznych), można oczekiwać mniejszej liczby ścieżek niepoprawnie wyspecyfikowanych i odwrotnie, czyli jeśli model jest źle wyspecyfikowany, powinna rosnąć liczba konkluzji o nieprawidłowej specyfikacji. Dlatego w poniższej pracy zostaną wykorzystane procentowe rozkłady otrzymanych konkluzji dla ścieżek w obrębie danego modelu, czyli: *nm%* to procent ścieżek poprawnie wyspecyfikowanych (*not misspecified*), *m%* procent niepoprawnie wyspecyfikowanych i analogicznie dla poprawnie i niepoprawnie wyspecyfikowanych ścieżek w oparciu o parametr EPC: *EPC:m%, EPC:nm%*. Dodatkowo rozważony zostanie poziom procentowy ścieżek niekonkluzywnych *i%* oraz wskaźnik *total.m%*, czyli ogólny procent nieprawidłowej specyfikacji (wszystkich typów). Jako uzupełnienie obliczeń zostanie przyjęty także średni poziom surowej wartości indeksów modyfikacji dla całego modelu (MI_M), średni poziom surowej wartości *ECP* (ECP_M) i średni poziom surowej wartości *NCP* (NCP_M).

1.4.4. Konkluzje

Oczywiście różne współczynniki dopasowania mają swoje wady i zalety, czasem służą nieco innym celom, są na różne sposoby krytykowane albo sugeruje się, że powinny być odczytywane wraz z innymi współczynnikami (na przykład: Davcik, 2014; Markland, 2007; Millsap, 2007; Shi, Lee i Maydeu-Olivares, 2019; Steiger, 2007; Stone, 2021). Celem niniejszej pracy nie jest jednak dyskusja na temat współczynników samych w sobie. Przedmiotem pracy jest jedynie sprawdzenie, czy miary dopasowania mogą być dobrym sposobem rozstrzygania tego, ile realnie wymiarów posiada dany konstrukt i który model z kliku alternatywnych jest tym poprawnym. W związku z tym **drugie pytanie zadawane w niniejszej pracy to: (P2) które miary dopasowania i oceny modelu CFA pozwalają na najbardziej jednoznaczne wnioski dotyczące akceptacji bądź odrzucenia modelu w przypadku, gdy konstrukt jest jednowymiarowy, ale badacz nie jest pewny, ile ma wymiarów**.

Współczynniki oparte o indeksy modyfikacji są stosunkowo nowym podejściem w dziedzinie analiz CFA i były dotychczas wykorzystywane raczej do eksploracji modeli. Nie ma więc w literaturze na ich temat informacji, czy nadają się do rozróżniania modeli prawidłowych i nieprawidłowych oraz w jaki sposób to rozróżnienie miałoby następować. To kolejny obszar badań niniejszej pracy doktorskiej – w obrębie pytania badawczego P2: oprócz klasycznych i znanych miar dopasowania zostaną uwzględnione także rzadko stosowane miary oparte o indeksy modyfikacji.

1.5. Jedno- i wielowymiarowe modele teoretyczne: możliwości rozróżnienia

Obszarem zainteresowania niniejszej pracy są konstrukty jednowymiarowe (Bartholomew, Knott i Moustaki, 2011; Timmerman, Lorenzo-Seva i Ceulemans, 2018). Jednowymiarowość oznacza jeden czynnik latentny w modelu CFA (Bhattacherjee, 2012; Hill, 2011) lub w analizie EFA (Auerswald i Moshagen, 2019; Finch, 2020; Ledesma, Valero-Mora i Macbeth, 2015). W przypadku pomiaru cech latentnych za pomocą zbioru obserwowalnych wskaźników można ogólnie mówić o wątpliwościach co do ich jedno- lub wielowymiarowości (pośrednio porusza: Collins i Lanza, 2009). Głównym powodem tych wątpliwości jest zwykle niejasny obraz zmiennej latentnej wynikający na przykład z a) słabych zależności między zmiennymi obserwowalnymi (tu: pytaniami kwestionariusza), b) niewielkiej trafności

teoretycznej poszczególnych zmiennych obserwowalnych (tu: pytań jedynie w niewielkim stopniu dotyczących zjawiska, które chcemy badać) i c) zanieczyszczenia obrazu zmiennej latentnej poprzez różnego rodzaju błędy (np. losowe wpływy wynikające ze zróżnicowania indywidualnego lub o nieznanym charakterze) albo wpływ innej zmiennej latentnej (np. wskaźniki jednej zmiennej latentnej mogą być jednocześnie wskaźnikami innej zmiennej latentnej) (Gustafsson i Åberg-Bengtsson, 2010; Reisei in., 2016; Slocum-Gori i Zumbo, 2011; Ziegler i Hagemann, 2015).

Wymienione powyżej problemy utrudniają, a nawet często uniemożliwiają decyzję dotyczącą tego, jaką strukturę odkryto empirycznie. W praktyce, gdy badacze chcą sprawdzić z użyciem CFA, który wariant czynnikowy ich kwestionariusza jest najbliższy prawdzie, zwykle porównują współczynniki dopasowania kilku modeli i wybierają najlepszy pod względem współczynników. Literatura przedmiotu rekomenduje tu pewne reguły, jak interpretować współczynniki dopasowania, często określając progi rozdzielające „dobre" wyniki od „złych", czyli tzw. punkty odcięcia *(rule of tumb*) (np. Konarski, 2009; Putnick i Bornstein, 2016). Niemniej jednak, jak łatwo zauważyć w publikowanych pracach, różnice między poszczególnymi współczynnikami dopasowania są często minimalne (występują np. na drugim czy trzecim miejscu po przecinku).

Oto kilka przykładów tego problemu. Kwestionariusz Depresji, Lęku i Stresu (DASS-42; Lovibond i Lovibond, 1995w polskiej adaptacji Makara-Studzińska i in., 2022) porównuje wartości modelu „dobrego" (RMSEA = 0,052; SRMR = 0,047; GFI = 0,870; CFI = 0,911) i „złego" (RMSEA = 0,068; SRMR = 0,056; GFI = 0,797; CFI = 0,845), gdzie na przykład różnice RMSEA wynoszą 0,016, a największy przyrost (w obrębie CFI), zaledwie przekracza 7%. Kwestionariusz Samooceny SES Rosenberga (1965) w polskiej adaptacji (Łaguna i in., 2007), dość szeroko cytowanej (ponad 300 cytowań w samej bazie Scholar w styczniu 2025), porównuje model jedno- i dwuczynnikowy, pokazując różnice: RMSEA = 0,15; GFI = 0,83; NFI = 0,75; CFI = 0,76 vs RMSEA = 0,14; GFI = 0,85; NFI = 0,78; CFI = 0,79. Skala Wsparcia w Relacjach Intymnych (Ilska i Przybyła-Basista, 2020) podaje dla dwóch modeli dokładnie te same współczynniki RMSEA = 0,036 (0,031-0,042); CFI = 0,991; oraz TLI = 0,990; a wskazanie „lepszego" modelu następuje przez wskazanie różnic między wartościami χ^2: 470,23 vs 477,69 (lub jak można łatwo obliczyć miary χ^2/df 1,748 vs 1,762). Test Satysfakcji z Życia Riverside (Margolis i in., 2019) porównuje modele które mają współczynniki CFI = 0,999, TLI = 0,999, RMSEA = 0,066 vs CFI = 0,999, TLI = 0,994 RMSEA = 0,065. Porównanie współczynników modeli CFI = 0,96; TLI = 0,95; SRMR = 0,05; RMSEA = 0,03 vs CFI = 0,95; TLI = 0,95; SRMR = 0,05; RMSEA = 0,03 można znaleźć

w Wielowymiarowej Skali Samokontroli (MSCS; Nilsen i in., 2020). Kwestionariusz Komunikacji Małżeńskiej (Plopa, 2008), jedno z najpopularniejszych narzędzi w obrębie psychologicznej diagnozy małżeńskiej, porównuje w podręczniku metody model o parametrach RMSEA = 0,08; GFI = 0,82; AGFI = 0,79 z modelem o parametrach RMSEA = 0,09; GFI = 0,79; AGFI = 0,76.

Przykłady można oczywiście mnożyć. Na dodatek takie wyniki często są niejednoznaczne - jeden współczynnik dopasowania modelu pokazuje, że model jest lepszy, inny że jest gorszy (np: Havigerová, Dosedlová i Burešová, 2018; Lamers i in., 2011; Marella i in, 2010; Shapiro i in., 2018; Schmitt, Sass, Chappelle i Thompson, 2018; Reichenheim i in., 2011). Nie chodzi przy tym o krytykę podanych analiz, ale zilustrowanie praktyki badawczej i publikacyjnej, w której badacze są zmuszeni do podejmowania pewnych decyzji na podstawie niejasnych kryteriów i w kontekście dużej niepewności co do rzeczywistego funkcjonowania miar dopasowania w danych empirycznych. Warto zauważyć, że większość zaleceń określa przede wszystkim, czy pojedynczy model jest „dobry" lub „zły", ale w obrębie porównywania alternatywnych modeli literatura przedmiotu nie daje badaczom jednoznacznych odpowiedzi w postaci konkretnych reguł odcięcia.

W przypadku struktur jednowymiarowych badacze przyjmują, że taka struktura powinna mieć wyraźnie lepszy poziom dopasowania w analizie CFA niż złożona z tych samych itemów struktura z większą liczbą wymiarów (np.: Lima-Rodríguez, Lima-Serrano i Domínguez-Sánchez, 2015; Rytilä-Manninen i in., 2016; Strelhow, Sarriera i Casas, 2020). Obecnie literatura przedmiotu nie jest konkluzywna w kwestii tego, czy rzeczywiście współczynniki dopasowania różnią się istotnie pomiędzy modelami jedno- i wielowymiarowymi utworzonymi na tych samych danych o niewątpliwie jednowymiarowej strukturze (wypełnienie tej luki jest celem niniejszej pracy), a w konsekwencji współczesny stan wiedzy w tej dziedzinie nie pozwala jednoznacznie określić, czy istnieją jakieś jednoznaczne reguły odcięcia dla współczynników dopasowania rozróżniające w tej grupie modeli prawidłowe od nieprawidłowych, co wydaje się być istotną luką w wiedzy o modelowaniu konfirmacyjnym. Zatem **trzecie i ostatnie pytanie badawcze (P3) niniejszej pracy dotyczy tego, czy da się wskazać punkty odcięcia dla miar dopasowania modeli CFA rozstrzygające, czy model jest rzeczywiście jednowymiarowy czy wielowymiarowy**. We wnioskach tej pracy zostanie zaproponowanych także kilka praktycznie użytecznych rozwiązań.

2. PROBLEM BADAŃ WŁASNYCH

Powyżej zidentyfikowano istotne luki w literaturze przedmiotu dotyczące kryteriów decyzji, czy uzyskane wyniki empiryczne wspierają model jedno- czy wielowymiarowy. Niniejsza praca ma na celu uzupełnienie tych luk.

W szczególności, podsumowując powyższy wywód teoretyczny do postaci pytań badawczych – w obrębie znanych metod porównania modeli z użyciem klasycznych miar dopasowania nadal nie jest jednoznaczne,

(P1) czy rzeczywiście współczynniki dopasowania różnią się istotnie pomiędzy modelami jedno- i wielowymiarowymi utworzonymi na tych samych danych o jednowymiarowej strukturze.

W związku z tym, ponieważ zostanie przetestowana większość popularnych miar dopasowania, zarówno globalnych, jak i lokalnych, zadano pytanie o to

(P2), które miary dopasowania i oceny modelu CFA pozwalają na najbardziej jednoznaczne wnioski dotyczące akceptacji bądź odrzucenia modelu w przypadku, gdy konstrukt jest jednowymiarowy, ale badacz nie jest pewny ile ma on wymiarów.

Ponadto w niniejszej pracy zostanie podjęta próba określenia

(P3) czy da się wskazać punkty odcięcia dla miar dopasowania modeli CFA rozstrzygające, który model jest rzeczywiście jednowymiarowy.

W celu odpowiedzi na sformułowane pytania badawcze przeprowadzono dwa rozbudowane badania i dwie odpowiadające im analizy: badanie symulacyjne i badanie empiryczne. W pierwszym badaniu wygenerowano dane symulacyjne (31500 symulacji) i poddano je analizie. W drugim badaniu zebrano dane empiryczne (łącznie 5172 osoby badane) i poddano je analizom analogicznym jak w badaniu 1.

3. BADANIE 1 (DANE SYMULACYJNE)

3.1. Metoda badania symulacyjnego

3.1.1. Wprowadzenie

Badanie symulacyjne przeprowadzono w paradygmacie Monte Carlo (Dimov, 2008; Dunn i Shultis, 2011). W obrębie tego paradygmatu wykorzystuje się zazwyczaj zróżnicowane metody generowania dużej liczby sztucznych danych o zadanych parametrach (zgodnych z określonymi wartościami wejściowymi, regułami prawdopodobieństwa, relacjami między zmiennymi itd.) i na ich podstawie estymuje się inne wartości wynikające z analizy wygenerowanych prób (na przykład estymacji błędów standardowych czy skuteczności oceny normalności rozkładów) (Chalmers i Adkins, 2020; Harding i in., 2014; Ventura-León i in., 2023). Przygotowano w tym celu skrypt w języku R 4.3.1 (R Core Team, 2023). Skrypt wraz z wynikami analiz jest dostępny w projekcie OSF pod linkiem https://osf.io/y36hz

Innymi słowy, w badaniu symulacyjnym skrypt generował sztuczne zestawy danych o różnych parametrach – każdy taki zbiór danych miał odzwierciedlać zestaw pytań kwestionariusza psychometrycznego – a następnie na tych zbiorach danych wykonywane były analizy konfirmacyjne i wyliczane współczynniki dopasowania. Otrzymane miary dopasowania dla pojedynczego symulowanego zbioru danych tworzyły pojedynczy punkt danych (*data point*[6]). Poszczególne punkty danych wygenerowano z użyciem różnych warunków wejściowych symulacji, a wszystkie punkty danych razem tworzyły matrycę wyników (*result matrix*[7]). Poniżej zostanie szczegółowo przedstawiony proces generowania danych, uzyskanie z nich potrzebnych współczynników do postaci matrycy wyników oraz analiza statystyczna tej matrycy.

---

[6] Termin nie ma w języku polskim dobrego odpowiednika. W przypadku tradycyjnych badań empirycznych grupy ludzi *data point* w obrębie danych może być zbliżone terminologią do *obserwacji*, czyli zestawu wyników dla jednej osoby z informacją o warunkach badawczych lub zapisanych wartości poszczególnych zmiennych (np. grupa badawcza vs kontrolna, wiek, wynik testu reakcji itd.). Terminu *data point* można użyć, gdy chodzi o *obserwację* wyniku nieuzyskanego w badaniach empirycznych grupy ludzi, czyli tutaj serię zapisanych wartości poszczególnych warunków symulacji (m. in. poziom korelacji i rodzaj skali Likerta) oraz wyników dopasowania w symulowanym zbiorze danych (m. in. miar dopasowania RMSEA, GFI, TLI, CFI i tak dalej).

[7] Należy tu wyraźnie rozróżnić, że *zbiór danych* to wygenerowany w obrębie pojedynczej symulacji „sztuczny" zbiór odpowiedzi osób badanych na pytania kwestionariusza, czyli symulacja badania, na którym następuje próba dopasowania modelu teoretycznego do danych; następnie taki *zbiór danych* jest przekształcany do *punktu danych*, przy czym nie jest zapisywany sam zbiór, ale współczynniki dopasowania dla tego konkretnego zbioru; z kolei zbiór *punktów danych* z bardzo dużej liczby pojedynczych symulacji na *zbiorach danych* jest łączony w bazę danych nazywaną dla rozróżnienia *matrycą wyników*.

3.1.2. Przygotowanie symulacji

Jako warunki wejściowe poszczególnych symulacji (kolejnych zbiorów danych) ustalano (kontrolowano): liczbę itemów (pozycji) symulowanego kwestionariusza, wielkość próby, średni poziom korelacji między pozycjami, rodzaj rozkładu, typ skal Likerta oraz liczbę zmiennych latentnych (patrz: Tabela 1). Liczba itemów była ustalana na 10, 20, 30, 40 lub 50 (symulowały one mniej lub bardziej rozbudowane kwestionariusze), wielkość próby wahała się od 100 do 1000 osób w symulacji (w przyrostach co 100 osób), a średni poziom korelacji między pozycjami przyjmował wartości od 0,1 do 0,9 w przyrostach co 0,1. Kolejne kontrolowane warunki wejściowe dotyczyły różnych typów danych w obrębie symulowanego kwestionariusza. Większość badań symulacyjnych po prostu tworzy zbiory danych oparte o liczby rzeczywiste[8] (Goldfeld i Wujciak-Jens, 2020; Peng, 2016). W psychologii najczęściej nie stosuje się w obrębie skal psychometrycznych liczb rzeczywistych, tylko skale Likerta (Cheng et. al., 2021; Taherdoost, 2019), dlatego oprócz danych opartych o liczby rzeczywiste wykonano ich rzutowanie na różne skale Likerta – od skal dychotomicznych, do skal siedmiostopniowych. Następny warunek badawczy dotyczył rozkładów odpowiedzi na pytania kwestionariusza (zmiennych): symulowano rozkłady jednorodne i rozkłady zbliżone do rozkładu normalnego.

Tabela 1
Warunki wejściowe badania symulacyjnego

Warunek (czynnik)	Poziomy zmiennej	Skrót
Rozkład	jednorodny (U) lub normalny (N)	DIS
Średnia korelacja między pozycjami	0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; 0,9	MIC
Liczba itemów kwestionariusza	10, 20, 30, 40, 50	NOI
Wielkość próby	100, 200, 300, 400, 500, 600, 700, 800, 900, 1000	SMP
Typ skal Likerta	0, 2, 3, 4, 5, 6, 7	LST
Liczba zmiennych latentnych	1, 2, 3, 4, 5	NOF

Adnotacja: 0 dla typu skal oznacza liczby rzeczywiste, pozostałe wartości oznaczają długość skali Likerta; „skrót” oznacza skróty użyte na wykresach i w tabelach

[8] Funkcja genCorGen z pakietu SimStudy (Goldfeld & Wujciak-Jens, 2020) umożliwia generowanie zmiennych losowych o wartościach rzeczywistych, których zakres zależy od przyjętego rozkładu: w przypadku rozkładu jednostajnego wartości mieszczą się w przedziale [0,1], natomiast dla rozkładu normalnego przyjmują zakres [−5,5]. W celu uproszczenia prezentacji wyników w tabelach i na wykresach, oraz dla jednoznacznego odróżnienia tego sposobu generowania zmiennych od różnych typów skal Likerta, przyjęto oznaczenie tej kategorii za pomocą wartości 0.

W literaturze dotyczącej określania struktury danych pojawia się kwestia nie tyle dychotomicznego rozróżnienia na jedno- i wielowymiarowe struktury, ile poziomu pewności, że dany konstrukt jest jednowymiarowy (Yuan i in., 2004) w przypadku danych o nieznanej wymiarowości. Wymienione wcześniej popularne metody eksploracyjne (na przykład analizy PCA/EFA), których zadaniem jest właśnie określenie liczby wymiarów i/lub ich redukcja, zakładają u podstaw, że nawet jeśli część problemów dotyczących ekstrakcji czynników (procesu, w którym decyduje się o potencjalnej liczbie wymiarów zbioru danych) będzie z ich użyciem nierozstrzygalna (Rayi in., 2021; Wangi in., 2017), to jednocześnie istnieje grupa zbiorów danych, która ma w swoim obrębie zmienne latentne o niewątpliwej strukturze (pozwalającej na ekstrakcję zmiennej latentnej ze zmiennych obserwowanych i w miarę jednoznaczną decyzję co do wymiarowości danej struktury) (Cunningham i Ghahramani, 2015; Khattree i Naik, 2000; Nguyen i Holmes, 2019). Podsumowując: ze świadomością istnienia grupy takich zbiorów danych, dla których ocena wymiarowości jest trudna lub niemożliwa, w niniejszej pracy wykorzystane metody obliczeniowe zastosowano do struktur o jednoznacznie jednowymiarowym charakterze (potwierdzenie jednowymiarowości można znaleźć w rozdziale poniżej), co do którego można mieć dość dużą dozę pewności[9].

Do wszystkich symulowanych zbiorów danych (sztucznych wyników kwestionariusza), zawierających jednowymiarowe struktury czynnikowe, utworzono zarówno poprawny, jednowymiarowy, jak i niepoprawny wielowymiarowy model konfirmacyjny. W przypadku modeli niepoprawnych (zawierających od dwóch do pięciu wymiarów) poszczególne itemy zostały przypisane do jednego z niepoprawnie wyróżnionych wymiarów. Przyporządkowanie polegało na równomiernym rozlosowaniu poszczególnych itemów (pytań w zbiorze danych) do ustalonej liczby konstruktów. Na przykład w zbiorze danych zawierającym 20 pytań (budujących w rzeczywistości jeden konstrukt) przyjęto błędny model dwóch cech latentnych i losowo po 10 pytań zostało przyporządkowanych do każdej z nich w testowanym modelu. W sytuacji, gdy liczba itemów (pytań kwestionariusza) nie była podzielna bez reszty przez liczbę czynników (zmiennych latentnych), jak w przypadku 30 pytań i czterech czynników, przypisanie odbywało się w sposób możliwie równomierny. Innymi słowy ustalano liczbę od 2 do 5 czynników latentnych (wymiarów kwestionariusza) dla jednowymiarowego zbioru danych, tworząc w ten sposób alternatywny (niepoprawny) model wielowymiarowy, mimo że struktura pytań była w istocie jednoczynnikowa. Miało to odpowiadać sytuacji, w której badacz nie wiedząc, że struktura operacjonalizowanego

[9] Można też powiedzieć, że zanieczyszczenie testowanych przeze mnie struktur jednowymiarowych jest pomijalne dla ogólnych problemów poruszanych w tej pracy.

konstruktu jest jednowymiarowa, próbuje potwierdzić, że istnieje kilka wymiarów. Oczywiście poprawna specyfikacja dotyczyła tylko modelu jednoczynnikowego, a specyfikacja wszystkich pozostałych była nieprawidłowa.

Powyższy zestaw warunków (rozkład zmiennych, rodzaj skal w kwestionariuszu, wielkość próby, liczba zmiennych (itemów), liczba czynników (zmiennych latentnych), poziom korelacji) dawał 2 x 7 x 10 x 5 x 5 x 9 = 31 500 pojedynczych punktów danych (Tabela 1) tworzących matrycę wyników. Dla każdego wygenerowanego zbioru danych były obliczane miary dopasowania CFA dla modelu jednoczynnikowego (prawidłowego) oraz wielowymiarowego (nieprawidłowego, zależnego od warunku symulacji), a te z kolei były zapisywane jako pojedynczy punkt danych. Wyniki obliczeń dla każdego pojedynczego punktu danych zawierały wszystkie wymienione wcześniej grupy miar dopasowania.

Podstawą symulacji był warunek jednowymiarowości symulowanego konstruktu, czyli jednego czynnika latentnego w obrębie wygenerowania zbioru danych (symulacji kwestionariusza psychometrycznego). Jeśli chodzi o sam algorytm leżący u podstaw symulacji, można założyć jego poprawne działanie przy tworzeniu tablic (zbiorów danych) jednowymiarowych. Funkcja *genCorGen* z pakietu *SimStudy* (Goldfeld i Wujciak-Jens, 2020), według dokumentacji technicznej (dostęp w repozytorium CRAN dla wersji 0.8.1 pakietu), generuje dane przy użyciu metody mnożenia ustalonej macierzy korelacji przez wektory losowe pobrane z zadanego typu rozkładu (na przykład normalnego lub jednorodnego). W skrócie zatem, proces generowania danych oparty jest o jedną zmienną przekształcaną z użyciem zadanej macierzy, a zatem są to po prostu zanieczyszczone losowym szumem transformacje jednej i tej samej zmiennej i z matematycznego punktu widzenia można uznać, że w rezultacie macierze tak utworzonych danych reprezentują jeden czynnik latentny wyrażony poprzez serię zmiennych (pośrednio: Bartholomew i in., 2011). Mimo tego, że podobne założenie jest z powodzeniem stosowane w badaniach symulacyjnych z wykorzystaniem różnych pakietów (Fialkowski i Tiwari, 2019; Koliopanos i in., 2023), w ramach niniejszej pracy zdecydowano, aby zweryfikować poprawność tego założenia. Dla generowanych zbiorów danych użyto więc metod określających ile czynników realnie miała symulowana struktura. Sprawdzenie odbywało się za pomocą analiz eksploracyjnych opartych o metody PCA / EFA, a rezultaty można traktować jako analizę kontrolną.

3.1.3. Kontrola warunków symulacji

Warunki wejściowe symulacji były ustalane tak, by w obrębie jednego zbioru danych poprawna była tylko struktura jednoczynnikowa (jedna zmienna latentna lub jeden wymiar

kwestionariusza). Jest wiele metod określających, ile wymiarów opisuje dany zbiór danych (Garson, 2022; Silva i in., 2020) i należy zwrócić uwagę, że niniejsza praca nie wyczerpuje wszystkich dostępnych opcji, ani nie rozstrzyga która z dróg decyzyjnych jest najlepsza (przegląd metod znajduje się w pracy Auerswald i Moshagen, 2019). Z dostępnych metod zostały wybrane te które: a) pozwalają na automatyzację, tzn. istnieją jednoznaczne wskazówki co do reguł decyzyjnych na temat liczby wymiarów, a decyzja podejmowana jest bez nadzoru badacza; b) były one dostępne w obrębie popularnych pakietów dla R, i c) ich czas pracy w obrębie symulacji był relatywnie krótki. Ten ostatni warunek był o tyle istotny, że rozmiary całej symulacji wymagały często tygodniowych obliczeń, a każde dodatkowe „polecenie" (na przykład dodanie analizy wykresu osypiska) wykładniczo zwiększało czas pracy oprogramowania (z estymowanych dni do miesięcy). Z tej przyczyny należało wykluczyć dość obciążające obliczeniowo pakiety dla metody Hull'a (Lorenzo-Sevai in., 2011) i metody EKC (Braeken i van Assen, 2017). Z powodu braku możliwości pełnej automatyzacji wykluczono też metodę Kaisera-Guttmana opartą o wartości własne (*eigenvalues*); dodatkowym powodem była także współczesna krytyka tej metody wskazująca na jej liczne wady (Lance i in., 2006).

W konsekwencji wybrano: test wykresu osypiska (kryterium Cattela, *scree plot*), analizę równoległą w wersji skorygowanej (*revised parallel analysis*), metodę Velicera w wersji skorygowanej (*minimum average partial (MAP) test revised*) oraz, jako punkt odniesienia, sekwencyjny test $\chi 2$ (*sequential chi-square model test*). Te cztery metody (opisane szerzej w Auerswald i Moshagen, 2019), można jednak uznać za wystarczające do rozstrzygnięcia, czy zadane w symulacji dane jako jednowymiarowe, procedura symulacyjna wygenerowała jako rzeczywiście jednowymiarowe.

Test wykresu osypiska bywa automatyzowany na potrzeby analiz w paradygmacie Monte Carlo i zwykle automatyzacja tego kryterium, *de facto* wizualnego (zob. Ledesma i in., 2015), polega na algorytmie analizującym kąty kolejnych połączeń między punktami wykresu, i tam gdzie kąt jest mniejszy niż 45 stopni, tam przyjmuje się, że kończy się „stromość" wykresu, zatem można mówić o automatycznym podjęciu decyzji w oparciu o kryterium wizualne. Analiza równoległa w wersji skorygowanej jest kryterium zarówno wizualnym jak i regułą odcięcia. W obrębie samej metody, po obliczeniu zachowanych skorygowanych wartości własnych (*retained adjusted eigenvalues*), przyjmuje się jako kryterium odcięcia wartość własną większą niż 1. Niemniej jednak, podobnie jak w przypadku wykresu osypiska, generowany jest wykres, w oparciu o który badacz może podjąć arbitralnie decyzję o tym, gdzie kończy się „stromość". Właśnie taki warunek został przetestowany, tzn. zautomatyzowano detekcję miejsca „załamania" wykresu wartości własnych z użyciem wyniku

analizy równoległej. Sekwencyjny test χ2 iteracyjnie dopasowuje modele z rosnącą liczbą czynników, oceniając ich dopasowanie do danych, a proces kończy się, gdy dodanie kolejnego czynnika nie poprawia istotnie modelu. Jednak metoda ta może zarówno przeszacowywać, jak i niedoszacowywać liczbę czynników, a przy dużych próbach szybko uzyskuje istotne wartości, nawet dla danych bez spójnej struktury (Auerswald i Moshagen, 2019). Z tego względu, mimo popularności metody, test ten pełnił jedynie funkcję punktu odniesienia w niniejszej analizie.

W przeważającej większości kryteria dawały satysfakcjonujący wynik (Tabela 2). Odsetek modeli, które nie zostały uznane za jednowymiarowe w przypadku analizy automatycznej wykresów osypiska (zarówno kryterium Cattela, jak i wykresu osypiska analizy równoległej) nie przekraczał 0,5%. W przypadku testu MAP i obu kryteriów wizualnych (kryterium Cattela i osypisko analizy równoległej) było to mniej niż 1,5%, a w przypadku klasycznej analizy równoległej poniżej 2,5%, co wydaje się być wielkością bez wpływu na ogólne wyniki uzyskanych analiz. Niski poziom rozpoznania modeli jako jednowymiarowe wskazała metoda sekwencyjnego testu χ2 (poziom wskazania symulowanych struktur jako jednowymiarowe był ledwo bliski 70%), ale biorąc pod uwagę niską wartość diagnostyczną tego testu i spójność pozostałych czterech wskaźników, wynik ten nie powinien być szczególnie brany pod uwagę. Na tym etapie można powiedzieć, że symulowane dane z bardzo dużą dozą pewności miały jeden ogólny czynnik latentny (strukturę jednowymiarową).

Jeśli chodzi o modele, które nie zostały rozpoznane jako jednowymiarowe, ich przeciętna korelacja między pozycjami była bardzo niska a zbiór danych zawierał relatywnie niewielką liczbę obserwacji (zob. Tabela 3, wraz z liczbą obserwacji wzrasta odsetek rozpoznań zgodnych z założeniem). Tym samym ten niewielki zbiór modeli może po prostu nie być możliwy do oceny z użyciem wybranych wyżej metod, ale nie oznacza to niewątpliwej wielowymiarowej struktury. Można je natomiast uznać za nierozstrzygalne. Te modele pozostawiono jednak w symulacji z uwagi na zgodność wszystkich miar wskazujących na jednowymiarowość zbiorów danych dla niemal wszystkich pozostałych ich typów oraz przytoczone wcześniej publikacje wskazujące na poprawność generowania danych jednowymiarowych w obrębie algorytmu.

Dodatkowo zaprezentowano przeciętne (uśrednione) wykresy osypiska i wykresy analizy równoległej dla modeli rozpoznanych jako jednowymiarowe (rozstrzygalne, około 99%) i modele nierozpoznane jako jednowymiarowe (nierozstrzygalne, około 1%) (Rysunek 1). Uśredniając trzy pierwsze położenia punktów wyznaczających wykresy osypiska (zarówno klasycznego wykresu osypiska jak i wykresu osypiska opartego na analizie równoległej) dla wszystkich analizowanych 31500 zbiorów danych można wyraźnie zobaczyć, że w przypadku

modeli o strukturze jednowymiarowej jest widoczny jeden silny czynnik i duża stromość wykresu.

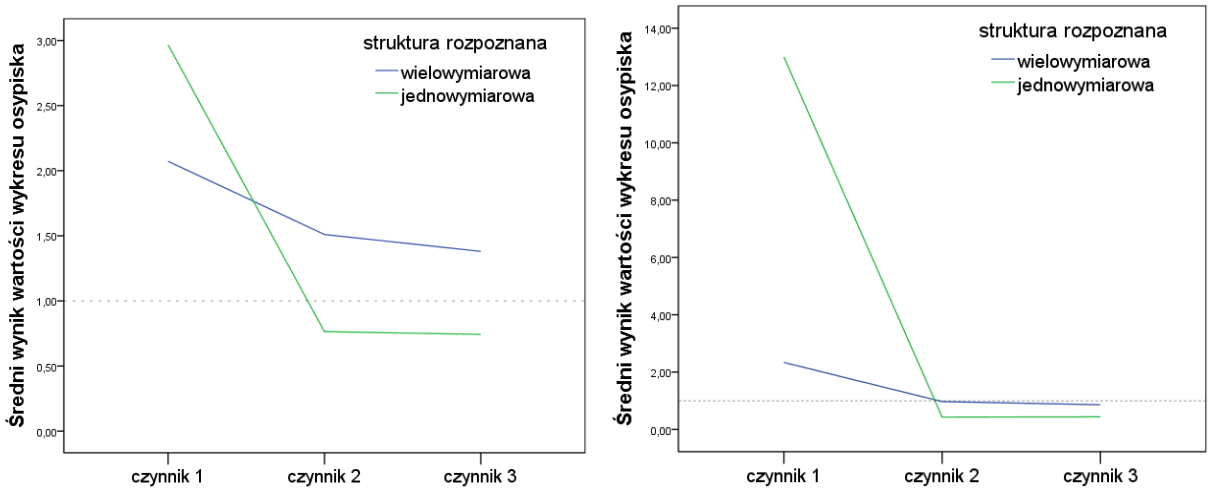
Tabela 2
Poziom rozpoznania symulowanych modeli jako jedno- i wielowymiarowych z użyciem różnych metod

	Model wielowymiarowy (niepoprawny)		Model jednowymiarowy (poprawny)	
	N	%	N	%
Kryterium Cattela	95	0,3%	31289	99,7%
Analiza równoległa	767	2,4%	30733	97,6%
Osypisko analizy równoległej	65	0,2%	31424	99,8%
Test MAP	398	1,3%	30365	98,7%
Sekwencyjny chi-^{2}	9957	31,8%	21400	68,2%

Tabela 3
Procentowy poziom rozpoznania modeli jako jednowymiarowe wraz ze wzrostem wielkości próby

Wielkość próby	Kryterium Cattela	Analiza równoległa	Osypisko analizy równoległej	Test MAP	Sekwencyjny chi-^{2}
100	98%	89%	99%	89%	67%
200	99%	94%	99%	99%	68%
300	100%	97%	100%	100%	68%
400	100%	98%	100%	100%	69%
500	100%	99%	100%	100%	68%
600	100%	99%	100%	100%	69%
700	100%	100%	100%	100%	68%
800	100%	100%	100%	100%	68%
900	100%	100%	100%	100%	68%
1000	100%	100%	100%	100%	68%

Rysunek 1. Przeciętne (uśrednione) wykresy osypiska (z lewej) i wykresy analizy równoległej (z prawej) dla modeli uznanych za jednowymiarowe (powyżej 99% obserwacji na wykresie) i wielowymiarowe (poniżej 1% wszystkich symulacji)



Adnotacja: Uśredniono wartości trzech pierwszych punktów (czynników) uzyskiwanych w analizie; dalsze czynniki w obu przypadkach były pomijalne.

3.1.4. Warunki brzegowe interpretacji miar dopasowania

Jako punkt odniesienia dla uzyskanych wyników miar dopasowania przyjęto wartości odcięcia (*cutoff, rule of thumb*) współczynników dopasowania, które uznaje się obecnie za satysfakcjonujące. Należy zwrócić uwagę, że w literaturze przedmiotu można znaleźć zalecenia zarówno mniej, jak i bardziej restrykcyjne (na przykład: Konarski, 2009; McNeish i Hancock, 2018; Sivo i in., 2006; Schermelleh-Engel, Moosbrugger i Müller, 2003; West, Wu, McNeish i Savord, 2023). Biorąc pod uwagę to zróżnicowanie w pracy przyjęto wartości dwojakiego charakteru. Tam, gdzie wraz z wielkością próby wymagane są wyższe poziomy miary dopasowania (ponieważ miara szybko przyrasta; Hu i Bentler, 1999), tam przyjęto bardziej konserwatywne punkty odniesienia. W przypadku pozostałych miar przyjęto po prostu poziomy współcześnie często spotykane w publikacjach psychologicznych i akceptowalne w praktyce jako poprawne dopasowanie. Wszystkie wartości odcięcia zostały zestawione w Tabela 4.

Do porównania uzyskanych wyników symulacji (a konkretniej współczynników dopasowania) modeli jedno- i wielowymiarowych użyto serii porównań międzygrupowych z użyciem testu rang U Manna-Whitney (MacFarland i in., 2016), ponieważ grupy były nierównoliczne, a rozkłady silnie odstające od rozkładu normalnego. Jako miarę siły efektu przyjęto miarę współczynnika korelacji dwuseryjnej Glassa (de la Rubia, 2022). W celach praktycznych, aby pokazać, gdzie różnice są bardzo małe, przedstawiono w tabelach zarówno

mediany, jak i średnie. Jako potencjalne punkty odcięcia obliczono współczynniki krzywej ROC (Pintea i Moldovan, 2009).

Aby ocenić, czy uzyskane wyniki nie są przypadkowym błędem pojedynczego algorytmu symulacji (błędu polegającego na zawieszeniu się którejś z pętli kodu albo bardzo specyficznego/niepowtarzalnego układu wylosowanych zmiennych), wygenerowano kilkadziesiąt matryc wyników. Analiza różnic w zestawach danych o tak dużym rozmiarze nastręcza pewnych trudności, ponieważ wszystkie, nawet najmniejsze różnice i najsłabsze efekty mogą uzyskiwać bardzo małe wartości *p*-istotności. Porównywano zatem stopień pokrycia się krzywych prezentowanych na dalszych wykresach, a więc oceny dokonano w zasadzie metodą wizualną. Ponieważ nie stwierdzono różnic pomiędzy nimi, dlatego można uznać, że przedstawiona tu matryca wyników jest typową. Kod w języku R służący do wygenerowania matrycy wyników oraz przykładowa matryca uzyskanych wyników znajdują się w repozytorium tej pracy pod adresem: https://osf.io/y36hz

3.2. Wyniki badania symulacyjnego

3.2.1. Wprowadzenie

Wyniki badania symulacyjnego zostaną przedstawione zgodnie z kolejnością analizowanych zagadnień badawczych. W pierwszej kolejności oceniona zostanie istotność różnic we współczynnikach dopasowania między modelami jedno- i wielowymiarowymi opartymi na tych samych danych o jednoznacznie jednowymiarowej strukturze (P1). Następnie przeanalizowane zostanie, które miary dopasowania i oceny modelu CFA umożliwiają najprecyzyjniejszą identyfikację rzeczywistej liczby wymiarów w konstruowanym kwestionariuszu (P2). W końcowym etapie podjęta zostanie próba wyznaczenia progów odcięcia dla miar dopasowania CFA, pozwalających jednoznacznie określić, czy analizowany model jest jedno- czy wielowymiarowy (P3).

Zostaną zaprezentowane i omówione tabele oraz grafiki wskazujące kluczowe zależności i różnice, aczkolwiek pominięte lub skrótowo opracowane zostaną te statystyki, które były powtarzalne; podstawowe statystyki opisowe, związki między analizowanymi zmiennymi i przykładowe wykresy dotyczące (omówionej niżej) wielkości próby i rozkładu wyników znajdują się w repozytorium do tej pracy. Również wszystkie pozostałe możliwe wizualizacje zostały przeniesione do tego repozytorium (https://osf.io/y36hz) w formie

zestawów danych i kodu w języku R pozwalającego na wygenerowanie ich w wysokiej rozdzielczości.

Na podstawie wstępnego przeglądu poszczególnych modeli okazało się, że liczba itemów oraz rodzaj rozkładu (jednorodny czy normalny), były właściwie bez znaczenia dla wszystkich uzyskanych wyników. Na niektórych wykresach dotyczących uzyskanych wyników (na przykład: Rysunek 5, Rysunek 8, Rysunek 28 itp.) zaprezentowano także te dwa warunki symulacji żeby pokazać, że rozrzut wyników w obrębie tych zmiennych jest proporcjonalny (losowy) dla wszystkich typów modeli. Zwiększanie wielkości próby we wszystkich warunkach badania poprawiało jakość współczynników dopasowania, ponieważ wraz ze wzrostem próby łatwiej uzyskać dopasowanie modelu (Konarski, 2009), ale ostatecznie wykluczono je z prezentacji z dwóch powodów. Po pierwsze wszystkie miary dopasowania i tak przy najmniejszej próbie przekraczały progi akceptowania wskaźników, a więc niezależnie od tego, jaka była wielkość próby, można było przeciętnie uzyskać przynajmniej akceptowalne dopasowanie analizowanych struktur. Po drugie we wszystkich typach modeli poprawa jakości wyników była proporcjonalna i bardzo mała – czyli zmiana była nieznaczna, niezależna od pozostałych warunków badawczych, a przyrost wskaźników dopasowania modelu był mniej więcej jednakowy w obrębie każdego typu modeli, a więc bez praktycznego zastosowania w kwestii poszukiwania rozstrzygnięć co do wymiarowości badanego narzędzia. Graficzna prezentacja tej kwestii została przeniesiona do załączników pracy (rysunki: Rysunek 55, Rysunek 56, Rysunek 57, Rysunek 58), a w plikach załączonych do niniejszej pracy jest to po prostu dodatkowa zmienna bez znaczenia dla uzyskanych tu konkluzji. Nieznaczne różnice notowano dla niektórych miar z powodu typów skal (dychotomie były gorzej dopasowane niż analogiczne dłuższe skale Likerta), ale zdecydowanie są to problemy rzutowania zmiennych ilościowych na porządkowe – ponownie na wykresach widać, że jest to kwestia nie związana z pozostałymi warunkami (omówiono szerzej poniżej).

3.2.2. Różnice współczynników w strukturach jedno- i wielowymiarowych

Pierwsze pytanie badawcze dotyczyło tego, (P1) czy współczynniki dopasowania różnią się istotnie pomiędzy modelami jedno- i wielowymiarowymi utworzonymi na tych samych danych o jednowymiarowej strukturze. Aby odpowiedzieć na to pytanie przeprowadzono analizę różnic w przeciętnym poziomie wszystkich wskaźników między modelami prawidłowymi i nieprawidłowymi. W Tabela 4 znajdują się porównania uśrednionych miar dopasowania modeli jednowymiarowych (prawidłowych) oraz modeli

wielowymiarowych (nieprawidłowych). Wyniki zestawione w tej tabeli, w celu odpowiedzi na pierwsze pytanie badawcze, zostały podzielone na jedno- vs wielowymiarowe (dwu-, trzy, cztero- lub pięciowymiarowe łącznie), stąd czterokrotnie większa liczebność drugiej grupy.

Tabela 4
Porównanie przeciętnych miar dopasowania dla modeli jednowymiarowych (poprawnych) i modeli wielowymiarowych (niepoprawnych)

		Modele wielowymiarowe (n = 25200)				Modele jednowymiarowe (n = 6300)					
	ref	M	SD	Mdn	Q1-Q3	M	SD	Mdn	Q1-Q3	p	rg
χ^2	---	602,40	545,29	443,89	165,12-917,69	608,24	546,77	448,30	169,81-921,08	0,013	0,020
χ^2 / df	< 5,00	1,18	0,33	1,11	1,02-1,24	1,19	0,33	1,11	1,02-1,24	0,998	0,000
χ^2_p	> 0,05	0,221	0,274	0,086	0,001-0,377	0,220	0,275	0,086	0,001-0,369	0,894	0,001
CFI	> 0,950	0,976	0,056	0,994	0,979-0,999	0,976	0,056	0,994	0,980-0,999	0,572	0,005
TLI	> 0,950	0,977	0,124	0,993	0,977-0,999	0,979	0,106	0,993	0,978-0,999	0,637	0,004
GFI	> 0,950	0,926	0,074	0,949	0,911-0,973	0,924	0,074	0,948	0,910-0,972	0,027	0,018
AGFI	> 0,950	0,913	0,081	0,940	0,895-0,964	0,913	0,081	0,940	0,894-0,964	0,923	0,001
RNI	> 0,950	0,978	0,097	0,994	0,979-0,999	0,980	0,094	0,994	0,980-0,999	0,712	0,003
RMSEA	< 0,05	0,019	0,019	0,014	0,006-0,026	0,019	0,019	0,014	0,006-0,025	0,881	0,001
AIC	---	28133,93	28993,22	19785,41	8505,02-39806,61	28141,28	28992,19	19762,44	8537,82-40009,17	0,981	0,000
BIC	---	28401,45	29069,90	20022,97	8698,48-40191,38	28388,21	29067,23	19945,58	8686,44-40330,95	0,961	0,000
RMR	< 0,020	0,022	0,027	0,012	0,007-0,025	0,022	0,028	0,013	0,007-0,026	0,141	0,012
SRMR	<0,050	0,0266	0,0157	0,0236	0,0158-0,0330	0,0270	0,0157	0,0242	0,0162-0,0336	0,016	0,020
nm%	---	60,15	33,51	76,20	40,38-85,71	75,70	35,64	93,33	86,21-94,87	<0,001	0,551
m%	---	4,23	8,43	1,67	,22-5,22	1,02	2,25	0,00	0,00-0,00	<0,001	0,544
epc:nm%	---	9,43	14,00	5,37	3,30-8,89	5,61	4,40	5,31	3,34-6,92	<0,001	0,064
epc:m%	---	1,78	8,94	0,00	,00-,00	0,03	0,16	0,00	0,00-0,00	<0,001	0,154
i%	---	24,41	32,32	10,38	2,96-24,62	17,64	36,07	0,00	0,00-0,00	<0,001	0,527
total m%	---	6,01	12,80	2,15	,35-6,00	1,04	2,25	0,00	0,00-0,00	<0,001	0,577
MI_M	---	127,30	4175,83	1,27	1,05-2,28	1,14	0,30	1,07	,99-1,18	<0,001	0,398
ECP_M	---	0,13	5,88	0,00	-0,01-0,04	0,00	0,00	0,00	0,000-,00	<0,001	0,294
NCP_M	---	651,32	4834,78	39,97	9,61-140,07	877,69	6372,06	51,37	11,67-177,91	<0,001	0,064

Adnotacja: kolumna „ref" oznacza wybraną wartość referencyjną dla danej miary dopasowania; wartości *p* przedstawiono dla testu *U* Manna-Whitneya; *Mdn* – mediana; *Q1-Q3* – pierwszy i trzeci kwartyl; *rg* - współczynnik korelacji dwuseryjnej Glassa;

W każdym przypadku, zarówno dla modeli jednowymiarowych (poprawnych), jak i wielowymiarowych (niepoprawnych), średnie wartości indeksów dopasowania przekraczały progi pozwalające na uznanie struktury za dopasowaną, przy czym jedynie GFI i AGFI przekraczały progi bardziej liberalne (0,90), a wszystkie pozostałe przekraczały progi wyższe i bardziej konserwatywne. Oznacza to, że samo przekroczenie sugerowanego przez literaturę danego progu akceptacji nie pozwalało rozróżnić modeli poprawnie dopasowanych (jednowymiarowych) i modeli niepoprawnych (wielowymiarowych). Zatem wskaźniki na tle proponowanych w literaturze punktów odcięcia nie okazały się przydatne do oceny jednego z modeli jako lepszego. Należy oczywiście zauważyć że zakres wyników dla wszystkich miar dopasowania był zdecydowanie szerszy (zob. Tabela 10), a więc zdarzały się także modele o słabszych parametrach, ale a) były one zdecydowanie rzadsze w wygenerowanych strukturach o charakterze jednowymiarowym, b) dane o gorszych współczynnikach uzyskiwały porównywalnie niskie wskaźniki zarówno w modelach jedno- jak i wielowymiarowych.

Mimo przekroczenia progów akceptowalnego dopasowania modeli można naturalnie analizować różnice między przeciętnymi wartościami w obu grupach. Jeśli chodzi o standardowe (klasyczne) miary dopasowania modelu, to istotne statystycznie były różnice jedynie dla surowej wartości ogólnej modelu χ^2 i SRMR. Należy jednak zauważyć, że chociaż różnica była istotna statystycznie, to miary siły efektu były bardzo małe (0,02), a różnice były właściwie niezauważalne, biorąc pod uwagę wartość χ^2 (odpowiednio średnia dla modeli wielowymiarowych M_W = 602,40 vs średnia dla modeli jednowymiarowych M_J = 608,24), a dla SRMR rozróżnialne dopiero na czwartym miejscu po przecinku (M_W =0,0242 vs M_J =0,0236), co w obu przypadkach, biorąc pod uwagę rozrzut możliwych wyników, właściwie wyklucza znalezienie diagnostycznego punktu odcięcia wskazującego poprawną strukturę modelu. Okazuje się więc, że na podstawie SRMR i χ^2 nie ma możliwości w praktyce jednoznacznego (nielosowego) określenia, że model wielowymiarowy był niepoprawnie dopasowany, mimo że testy wskazują istotne między nimi różnice.

Istotne i wyraźne różnice (większe siły efektu) zostały zaobserwowane dla wszystkich współczynników opartych o indeksy modyfikacji. Wyniki wskazują przede wszystkim, że niepoprawnie wyspecyfikowane modele miały większy odsetek ścieżek, które są (wg metody Saris i in., 2009) niepoprawnie wyspecyfikowane (*misspecified*) i niekonkluzywne, oraz jednocześnie mniejszy odsetek ścieżek, które zostały uznane za nie-nieprawidłowo wyspecyfikowane (*not misspecified*). Niepoprawnie wyspecyfikowane modele miały też dużo wyższe wartości indeksów modyfikacji (*MI*), wyższy poziom *ECP* i niższy poziom *NCP*.

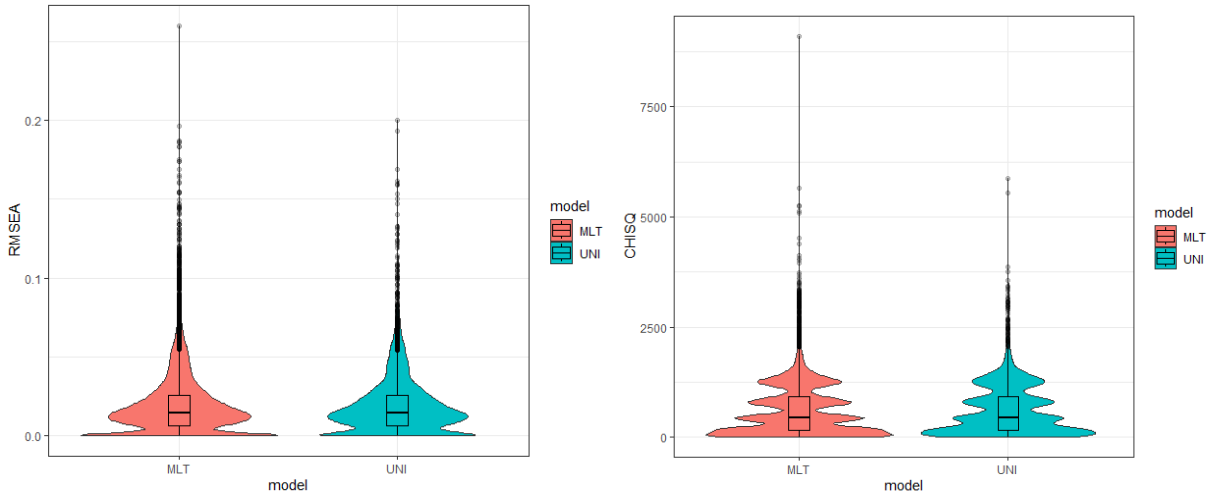
Udzielając zatem odpowiedzi na pytanie badawcze P1 należy stwierdzić, że popularne współczynniki dopasowania nie różnią się wyraźnie pomiędzy modelami jedno- i wielomiarowymi utworzonymi na tych samych danych o jednowymiarowej strukturze. Dlatego, w poszukiwaniu innych rozwiązań, w dalszej części przeprowadzono szczegółowy przegląd parametrów modeli, który jednak pokazuje, że potencjalne różnice między współczynnikami są właściwie pozorne.

3.2.3. Użyteczność miar dopasowania w rozstrzyganiu o wymiarowości

Drugie pytanie badawcze (P2) dotyczyło tego, które miary dopasowania i oceny modelu CFA pozwalają na najbardziej jednoznaczne wnioski dotyczące akceptacji bądź odrzucenia modelu w przypadku, gdy konstrukt jest jednowymiarowy, ale badacz nie jest pewny, ile ma wymiarów. Aby odpowiedzieć na to pytanie, przeanalizowano szczegółowo wykresy uwzględniające bardziej złożone zależności między zmiennymi, w szczególności to, jak wspólny wpływ różnych warunków badawczych (na przykład średniej korelacji między pozycjami i liczby czynników) może zmieniać poszczególne parametry dopasowania. Przed prezentacją tych wyników należy jednak poczynić dwa spostrzeżenia.

Pierwsza kwestia dotyczy faktu, że – jak wcześniej wspomniano – proste różnice między grupami mogą stanowić artefakt wynikający z wielkości próby, na co wskazują m.in. niskie wartości siły efektu. Aby wykazać, że istotność statystyczna różnic między porównywanymi współczynnikami nie jest szczególnie użyteczna w sytuacji, gdy badacz musi podjąć decyzję o poprawności modelu na podstawie ustalonych progów odcięcia, szczegółowej analizie poddano rozkłady wyników dla modeli jedno- i wielowymiarowych. Jako przykład najbardziej jaskrawy, przedstawione zostaną wyniki dla surowej wartości χ^2, która – w mniejszym lub większym stopniu – stanowi podstawę obliczeń większości współczynników dopasowania i dla której różnice między grupami okazały się istotne statystycznie, oraz dla równie podstawowej miary RMSEA, dla której nie odnotowano istotnych różnic (Rysunek 2).

Rysunek 2. Poziomy RMSEA i χ^2 dla modeli jedno- (UNI) i wielowymiarowych (MLT);

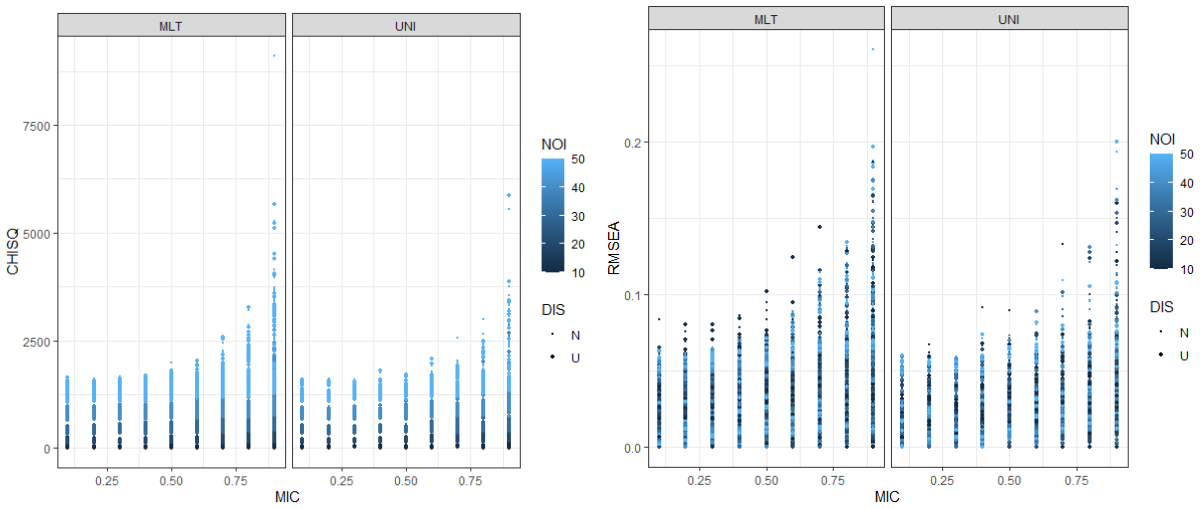


Adnotacja. Szerokość kolorowego obszaru wskazuje częstość występowania danego wyniku, wewnątrz znajdują się słupki średnich oraz błędu.

Na Rysunek 2 widać dokładniej, że niezależnie od tego, czy model był prawidłowy czy nieprawidłowy, właściwie z tym samym prawdopodobieństwem można było uzyskać wyniki świadczące o dopasowaniu. Tendencje centralne są praktycznie nierozróżnialne – przy czym warto przypomnieć, że dla wartości χ^2 różnice były istotne statystycznie, a dla RMSEA już nie. Ten przykład ilustruje, że ewentualne (istotne statystycznie) różnice międzygrupowe mogą wynikać jedynie z liczby obserwacji, a próba ustalenia na ich podstawie punktu odcięcia daje przypadkowe (nieprawidłowe) wyniki.

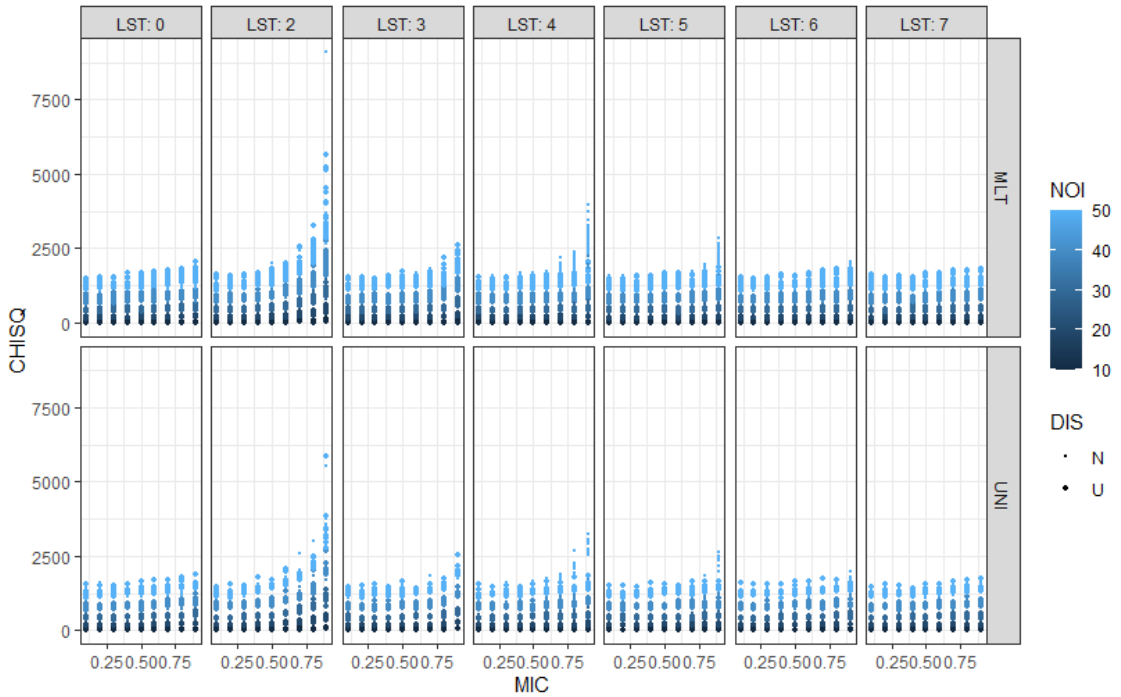
Druga uwaga dotyczy problemów rzutowania zmiennych ilościowych na skale dychotomiczne. Podwójny Rysunek 3 przedstawia rozrzuty uzyskanych wyników miar χ^2, a co i RMSEA (odpowiednio lewy i prawy obszar) dla modeli jedno- i wielowymiarowych (odpowiednio oznaczone skrótami UNI i MLT) w zależności od przeciętnego poziomu korelacji między pozycjami (oś *X*). Uwzględniono też na tym wykresie kolorem liczbę itemów (pytań kwestionariusza) a wielkością punktów rodzaj rozkładu (normalny vs jednorodny).

Rysunek 3. Pozorny przyrost χ^2 i RMSEA dla modeli jedno- (UNI) i wielowymiarowych (MLT) wraz ze wzrostem korelacji pomiędzy itemami (MIC); na wykresach zawarto też informację o liczbie itemów na konstrukt (NOI) i typie rozkładu (DIS)



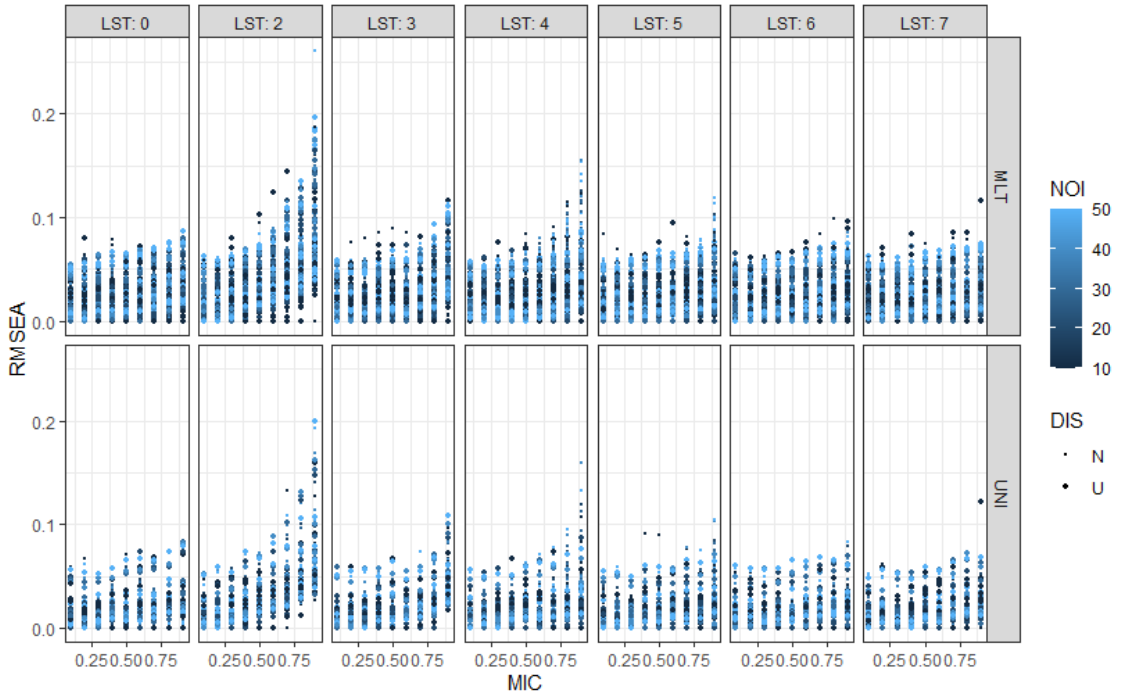
Jeśli brać pod uwagę jedynie średnią korelację między pozycjami (MIC) w obrębie danego konstruktu (Rysunek 3), może się wydawać na pierwszy rzut oka, że wraz z siłą korelacji rośnie wartość χ^2, a co za tym idzie i wartości RMSEA. Jednak wystarczy rozdzielić uzyskane wyniki względem typu wykorzystanej w badaniu skali, by okazało się, że problem ten dotyczy właściwie tylko skal dychotomicznych, które znacznie zawyżają wyniki współczynników przy dużych korelacjach. Poniżej Rysunek 4 oraz Rysunek 5 prezentują bardziej szczegółowe ujęcie, uwzględniając rodzaj skal Likerta użytych w generowaniu wyników oraz ich analizie (reprezentowane poprzez rozbicie wyników na kolejne pod-okna wykresu).

Rysunek 4. Pozorny przyrost χ2 dla modeli jedno- (UNI) i wielowymiarowych (MLT) wraz ze wzrostem korelacji pomiędzy itemami (MIC) z uwzględnieniem rodzaju skali (LST)



Adnotacja: Dla typu skali 0 oznacza liczby rzeczywiste, a wartości 2-7 różne typy skal Likerta, od dwustopniowej do siedmiostopniowej; na wykresie zawarto też informację o liczbie itemów na konstrukt (NOI) i typie rozkładu (DIS)

Rysunek 5. Pozorny przyrost RMSEA dla modeli jedno- (UNI) i wielowymiarowych (MLT) wraz ze wzrostem korelacji pomiędzy itemami (MIC) z uwzględnieniem rodzaju skali (LST)



Adnotacja: Dla typu skali 0 oznacza liczby rzeczywiste, a wartości 2-7 różne typy skal Likerta, od dwustopniowej do siedmiostopniowej; na wykresie zawarto też informację o liczbie itemów na konstrukt (NOI) i typie rozkładu (DIS)

W podrozdziale *3.1. Metoda badania symulacyjnego* zwrócono uwagę, że właściwie wszystkie dostępne obecnie generatory danych o wysokiej jakości posługują się głównie liczbami rzeczywistymi, które mogą zawierać dowolne wartości numeryczne (np. w obrębie symulowanego wyniku osoby badanej uzyskamy wartości 0,458 lub -4,907). Ponieważ badani zwykle udzielają odpowiedzi na skalach Likerta, gdzie są tylko wartości całkowite i to w konkretnych zakresach (na przykład tylko wartości 1, 2, 3 i 4), używa się - dla zwiększenia realizmu danych - rzutowania wartości rzeczywistych na wartości skal Likerta[10]. Oznacza to spore trudności w uzyskaniu rzeczywiście zróżnicowanych macierzy, gdzie dla skal dychotomicznych są tylko dwa możliwe rozwiązania (odpowiedzi osoby badanej). Biorąc pod uwagę zadaną wielkość korelacji między pozycjami na poziomie 0,8-0,9, kwestię dużej liczby itemów (40-50) oraz potrzebne na poziomie generatora i analizatora wyników zróżnicowanie w obrębie poszczególnych przypadków (symulowanych osób badanych), uzyskanie założonych korelacji jest niezwykle trudne, co powoduje po prostu zanieczyszczenie danych. Zanieczyszczenie danych prowadzi w efekcie do niższego dopasowania modelu oraz słabszych współczynników. Niemniej jednak, jak widać problem ten dotyczy tak naprawdę jedynie grupy modeli z dwustopniową skalą Likerta, a więc można uznać, że jest to jedynie problem rzutowania danych. Zatem przyrost wartości χ^2 oraz pogorszenie w jego efekcie współczynników dopasowania wraz z poziomem korelacji jest wynikiem zanieczyszczenia modeli dla skal dychotomicznych[11] i właściwie ma iluzoryczny charakter wynikający z ograniczeń metody obliczeniowej. Do kwestii tej wrócę jeszcze podczas przedstawiania poszczególnych wyników.

[10] Z pewnej perspektywy można to uznać za wadę generatorów, ale z drugiej strony rzadko się prowadzi badania na modelach strukturalnych w obrębie na przykład skal dychotomicznych, problem więc należy do otwartych.

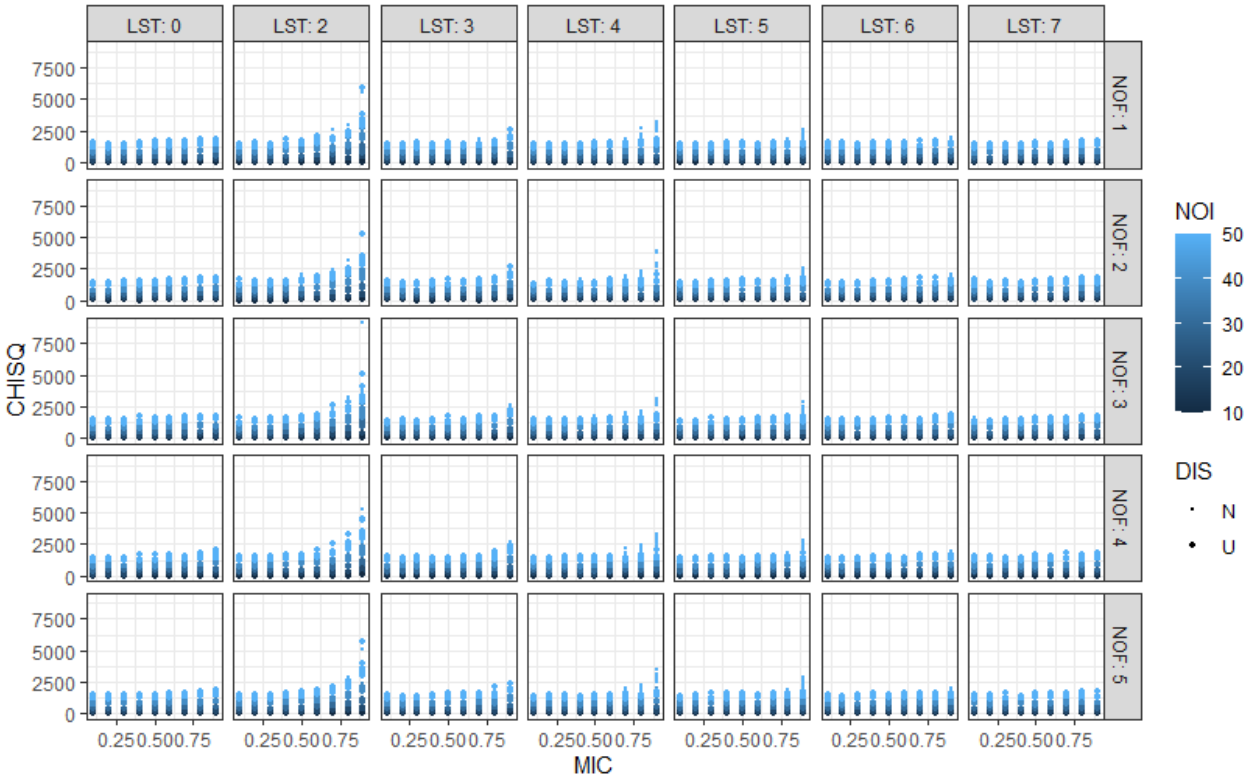
[11] Osobnym problemem wydaje się kwestia adekwatności metod SEM dla danych o charakterze małych skal, zwłaszcza dychotomii. To, które metody SEM byłyby rozwiązaniem problemów związanych z danymi o charakterze dychotomicznym, jest osobną kwestią, która obecnie nie wydaje się być rozstrzygnięta (porusza ten problem Song i Lee, 2005), ale nie jest to obszar zainteresowania niniejszej pracy

Na Rysunek 3, Rysunek 4 oraz Rysunek 5 widać jeszcze jedną kwestię wymagająca omówienia. Jak już wspomniano, ani rodzaj rozkładu (normalny vs jednorodny) ani liczba itemów nie miały wpływu na wyniki uzyskane dla wskaźników dopasowania. Na tych trzech i następnych wykresach widać wyraźnie, że rodzaj rozkładu, reprezentowany przez punkty o różnej wielkości, nie jest wyraźnie zróżnicowany w przestrzeni – punkty po prostu się całkowicie pokrywają. Co do liczby itemów (NOI, liczba pozycji kwestionariusza) w zasadzie tylko miara χ^2 wzrastała wraz z ich liczbą (ciemniejsze kolory punktów są położone zwykle niżej niż jasne kolory punktów), ale dotyczy to tylko miary χ^2, a wszystkie pozostałe wykresy prezentujące wyniki dla różnych liczb itemów są już losowo zróżnicowane. Ponadto proporcja przyrostu χ^2 jest zbliżona dla wszystkich typów wykresów, a więc bez znaczenia dla kwestii odróżniania modeli jedno- od wielowymiarowych. Dlatego będą to kwestie zasadniczo pomijane przy dalszej prezentacji wyników.

3.2.3.1. Miary globalne

Przejdźmy teraz do wyników szczegółowych dla wszystkich miar dopasowania zawartych w obrębie niniejszej pracy. Wynik szczegółowy oznacza tu prezentację poziomu danego współczynnika dopasowania w rozdziale na poszczególne warunki badawcze z uwzględnieniem średniej korelacji między pozycjami (MIC), typu skali (LST), liczby dopasowywanych czynników (NOF), liczby itemów (NOI) oraz rodzaju rozkładu (DIS). Na wykresie poniżej i na wszystkich następnych analogicznych wykresach szczegółowych zbiorczo przedstawiono serię wykresów rozrzutu (każde pojedyncze okno wykresu) pomiędzy średnim poziomem korelacji pomiędzy itemami (oś *X*) a wynikami dla analizowanego współczynnika (na przykład $\chi 2$ na Rysunku 6). Podział okien wykresów w poziomych wierszach jest związany z liczbą czynników w modelu, a w pionowych kolumnach są różne typy skal Likerta w symulacji: zatem okno w drugiej kolumnie (od lewej) i trzecim rzędzie (od góry) zawiera wyniki dla dychotomicznych skal Likerta i trzech czynników ustalonych w strukturze modelu SEM symulacji. Kolory punktów reprezentują liczbę itemów kwestionariusza - od ciemniejszych (10 itemów) do jaśniejszych (50 itemów), a wielkość punktów odpowiada typowi rozkładu symulacji (małe punkty to rozkład normalny, duże punkty to rozkład jednorodny).

Rysunek 6. Poziom χ^2 dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST) oraz różnej liczby czynników (NOF) i liczby itemów (NOI)

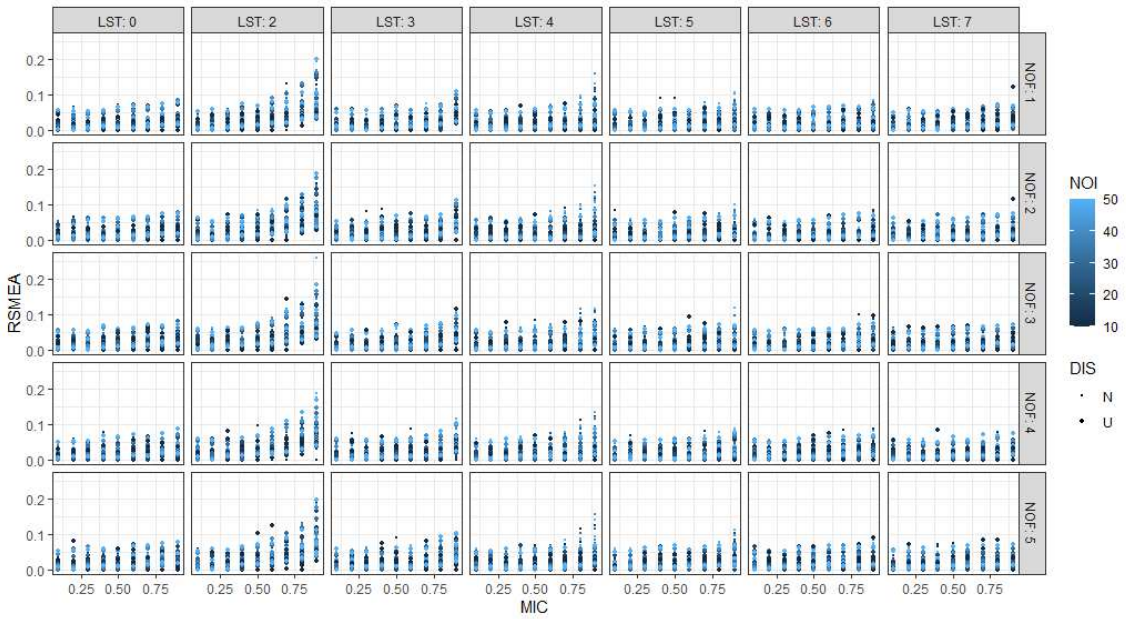


Poziom surowej wartości χ^2 (Rysunek 6), biorąc pod uwagę zakres wyników w każdym z okien wykresu w rozdziale na wszystkie wartości korelacji między pozycjami, okazał się zbliżony dla wszystkich typów modeli – niezależnie od tego, czy były to modele prawidłowe, czy nieprawidłowe. Każdy z modeli jednowymiarowych można było z takim samym prawdopodobieństwem dopasować jako dwu-, trzy-, cztero- lub pięciowymiarowy, a jedyne odstępstwa od przeciętnych wyników dotyczyły tak naprawdę wspomnianego problemu dla skal dychotomicznych. Analogiczny wynik uzyskano dla wartości χ^2/df (Rysunek 7) oraz dla poziomu istotności statystycznej wyniku surowego χ^2 (Rysunek 8); rozrzut wyników jest taki sam dla wszystkich warunków badania (pomijając oczywiście kwestię modeli opartych o skale dychotomiczne). Miary te powinny rozstrzygać o tym, czy model jest „dobry", ale w świetle przedstawianych tu wyników, z niemal takim samym prawdopodobieństwem można było uznać za dopasowany model prawidłowo wyspecyfikowany, jak i model nieprawidłowo wyspecyfikowany.

Rysunek 7. Wartość χ^2/df dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST) oraz różnej liczby czynników (NOF) i liczby itemów (NOI)

Rysunek 8. Wartość χ^2_p dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST) oraz różnej liczby czynników (NOF) i liczby itemów (NOI)

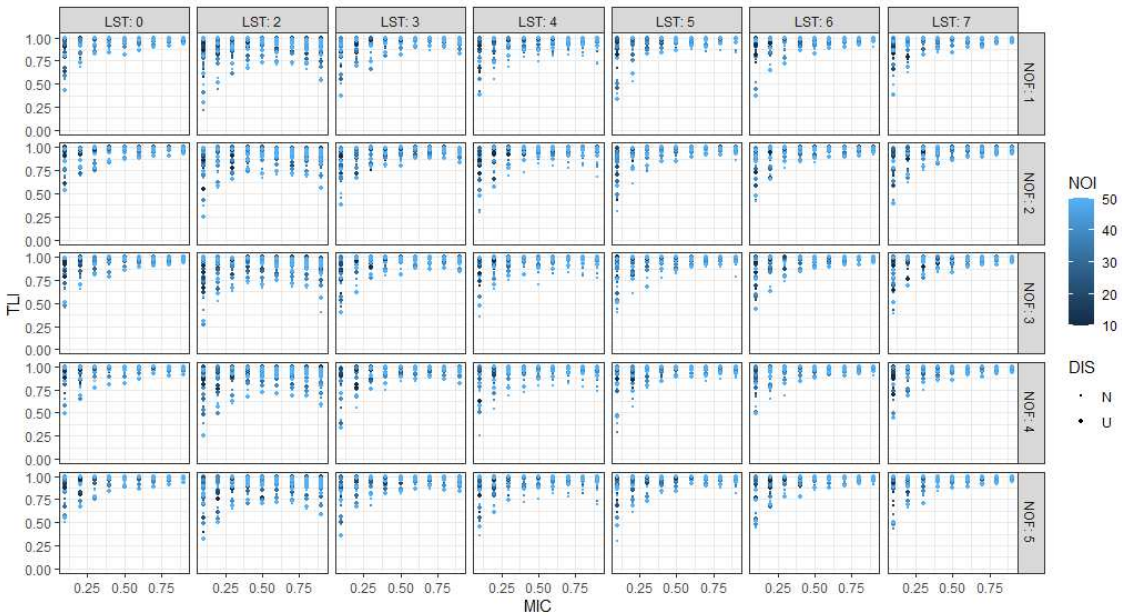
Rysunek 9. Poziom RMSEA dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST) oraz różnej liczby czynników (NOF) i liczby itemów (NOI)



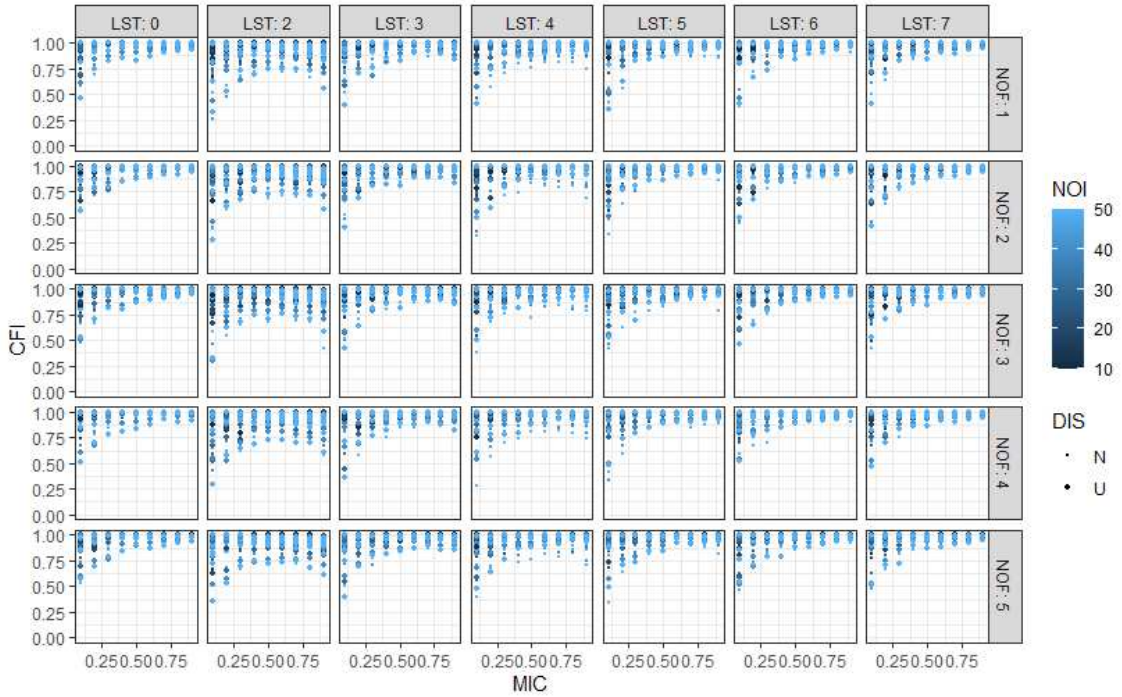
W przypadku RMSEA (Rysunek 9) dla prawie wszystkich modeli (prawidłowych i nieprawidłowych) miara ta utrzymywała się na poziomie akceptowalnym. Jedynie w obrębie skal o charakterze dychotomicznym, a także wysokich korelacji pomiędzy pozycjami, przyjmowała czasem wartości poniżej progu akceptowalności, co wynika ponownie ze wspomnianej liniowej metody rzutowania zmiennych o wartościach rzeczywistych na dychotomiczne skale. Oznacza to, że i ta miara nie nadaje się do rozróżnienia modeli poprawnych i niepoprawnych, bo wszystkie modele miały taką samą szansę uzyskać akceptowalne dopasowane.

Miara TLI (Rysunek 10) w większości przypadków plasowała się na poziomie satysfakcjonującym, przy czym widocznie lepsza była tam, gdzie średnie związki między pozycjami (itemami z których budowane były zmienne latentne) były wyższe i dawała gorsze rezultaty tam, gdzie związki między pozycjami były niższe. Co wyraźniej widać na Rysunek 10, można z użyciem tej miary określić mniej więcej, jak spójne są czynniki latentne w modelu (jak silnie poszczególne itemy związane są w obrębie czynnika latentnego), ale rozróżnienie pomiędzy modelami jedno- (poprawnymi) i wielowymiarowymi (niepoprawnymi) jest z jej użyciem niemożliwe. Podobne rezultaty można zaobserwować dla miary CFI (Rysunek 11) oraz miary RNI (Rysunek 12).

Rysunek 10. Poziom TLI dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST) oraz różnej liczby czynników (NOF) i liczby itemów (NOI)



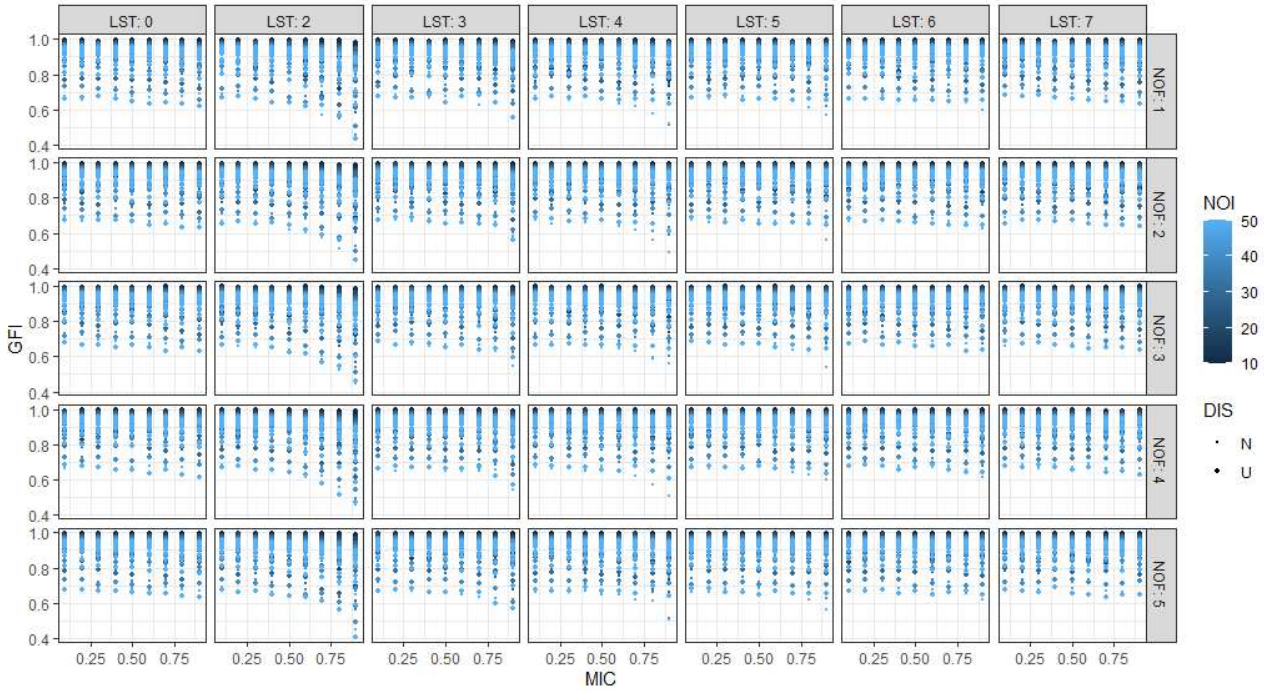
Rysunek 11. Poziom CFI dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST) oraz różnej liczby czynników (NOF) i liczby itemów (NOI)



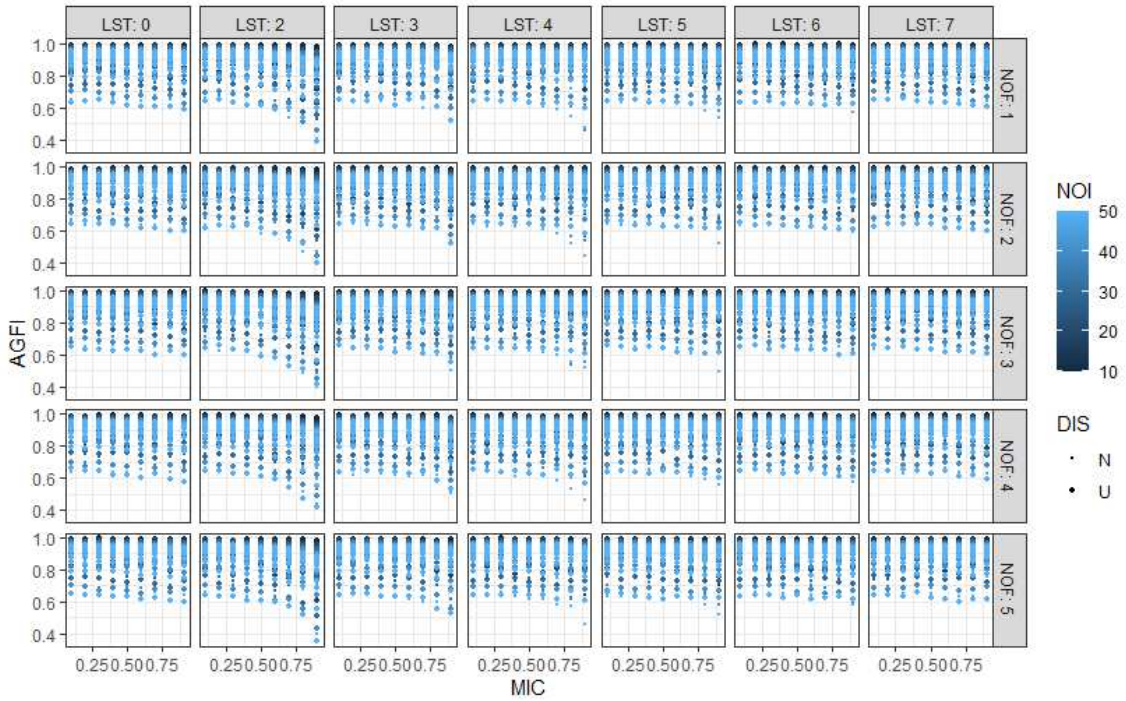
Rysunek 12. Poziom RNI dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST) oraz różnej liczby czynników (NOF) i liczby itemów (NOI)

Miary takie jak GFI i AGFI (Rysunek 13, Rysunek 14) były niemal tożsame (rs = 0,987; p < 0,001) i właściwie nierozróżnialne na poziomie poszczególnych klas modeli. Większość uzyskanych wskaźników GFI/AGFI sugerowało akceptowalne lub dobre dopasowanie, zarówno dla poprawnych jak i niepoprawnych modeli. Zdarzały się także pojedyncze modele o niskich wartościach tych miar, ale w każdym przypadku z takim samym prawdopodobieństwem dla wszystkich typów modeli. Ponownie wydają się to być wyniki o charakterze czysto losowym, a określenie na podstawie tych miar, który model (jedno czy wielowymiarowy) jest poprawny (lub lepszy), jest po prostu niewykonalne.

Rysunek 13. Poziom GFI dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST) oraz różnej liczby czynników (NOF) i liczby itemów (NOI)

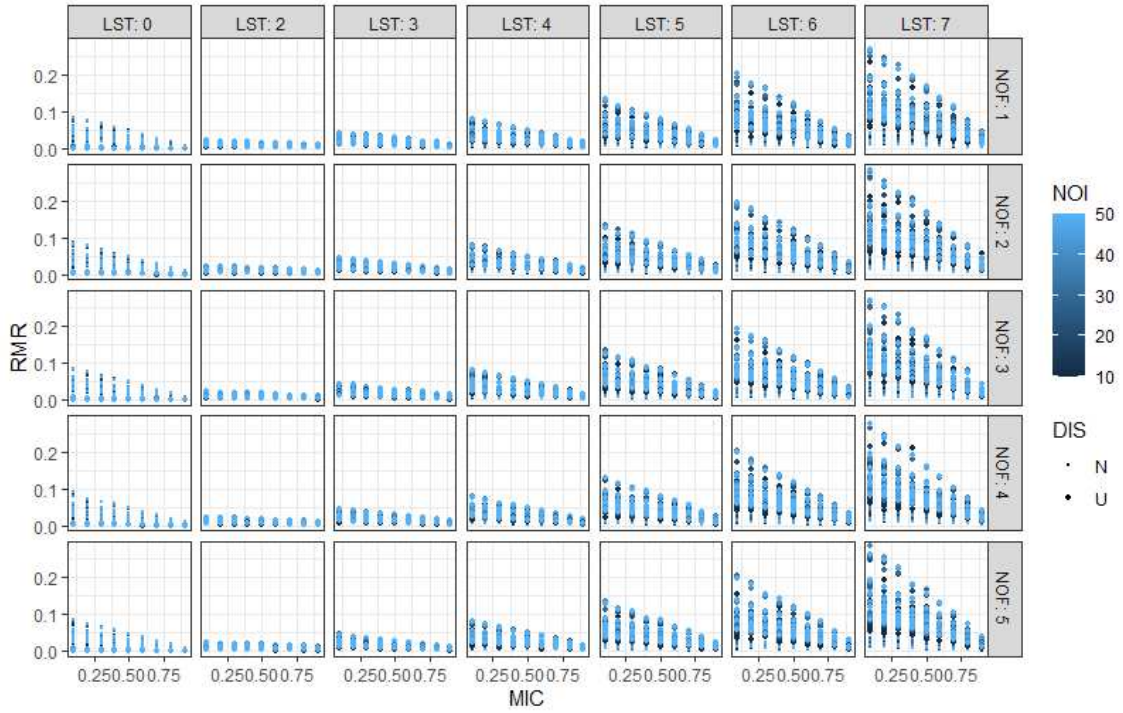


Rysunek 14. Poziom AGFI dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST) oraz różnej liczby czynników (NOF) i liczby itemów (NOI)

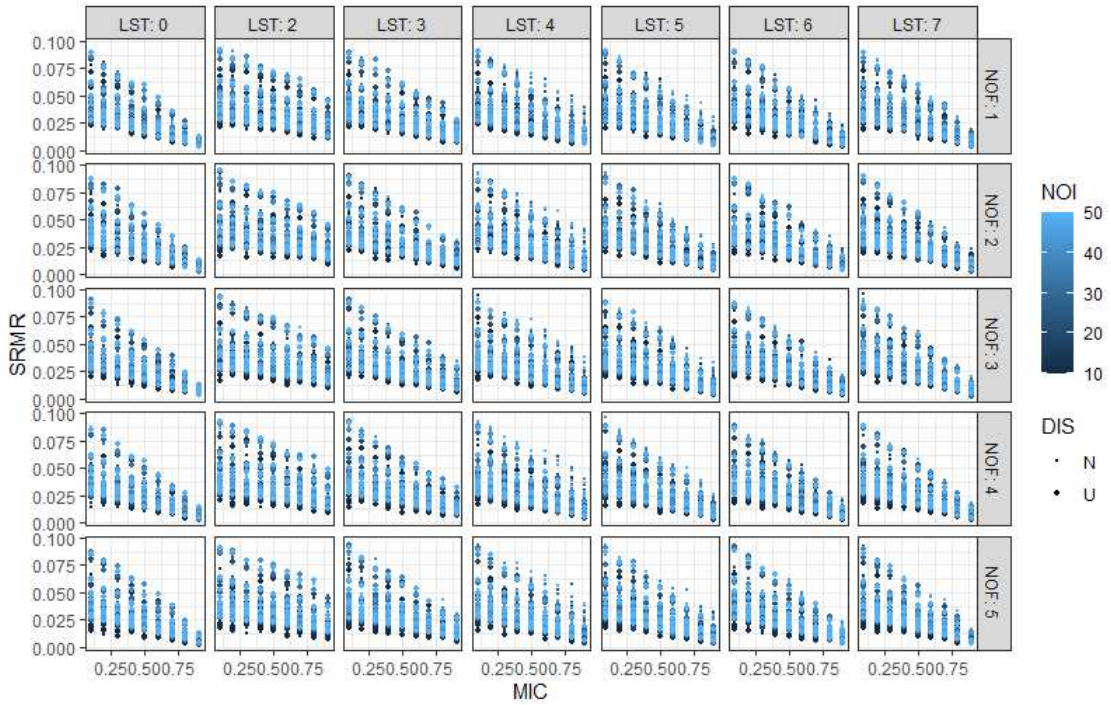


Interesujące układy wyników uzyskano dla RMR (Rysunek 15) i SRMR (Rysunek 16). W przypadku zarówno RMR i SRMR lepsze parametry dopasowania można uzyskać w przypadku wyższego poziomu korelacji, ale dla RMR zdecydowanie mniejsze odstępstwa (lepsze dopasowanie) dawały krótsze skale. W przypadku obu współczynników niestety bez znaczenia były pozostałe warunki badawcze, włącznie z kwestią jedno- i wielowymiarowości. Niezależnie od tego, ile zmiennych latentnych proponowano dla danej struktury, wyniki były zbieżne, a więc brak i tutaj możliwości rozróżnienia modeli prawidłowych od nieprawidłowych.

Rysunek 15. Poziom RMR dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST) oraz różnej liczby czynników (NOF) i liczby itemów (NOI)



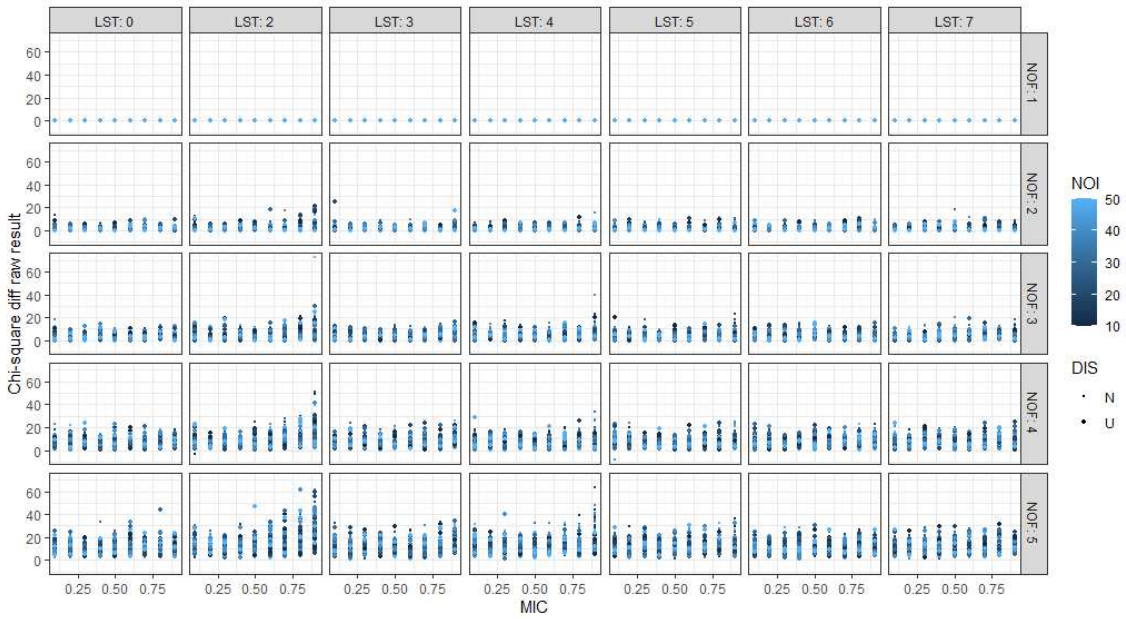
Rysunek 16. Poziom SRMR dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST) oraz różnej liczby czynników (NOF) i liczby itemów (NOI)



3.2.3.2. Miary przyrostowe

Teraz zostaną omówione popularne miary różnic między modelami. Jeśli chodzi o miarę χ^2_{diff} (Rysunek 17), która może określać różnicę między modelem jednowymiarowym (poprawnym) a modelem wielowymiarowym (niepoprawnym), to rzeczywiście rosła ona wraz z liczbą zmiennych latentnych w modelu. Jednak biorąc pod uwagę jej istotność statystyczną $\chi^2_{diff\text{-}p}$ (Rysunek 18) okazywało się, że rozrzut tych wartości jest tak duży, że różnice dla wszystkich modeli interpretowane tą miarą są właściwie losowe i można z takim samym prawdopodobieństwem uznać model za poprawny jak i niepoprawny.

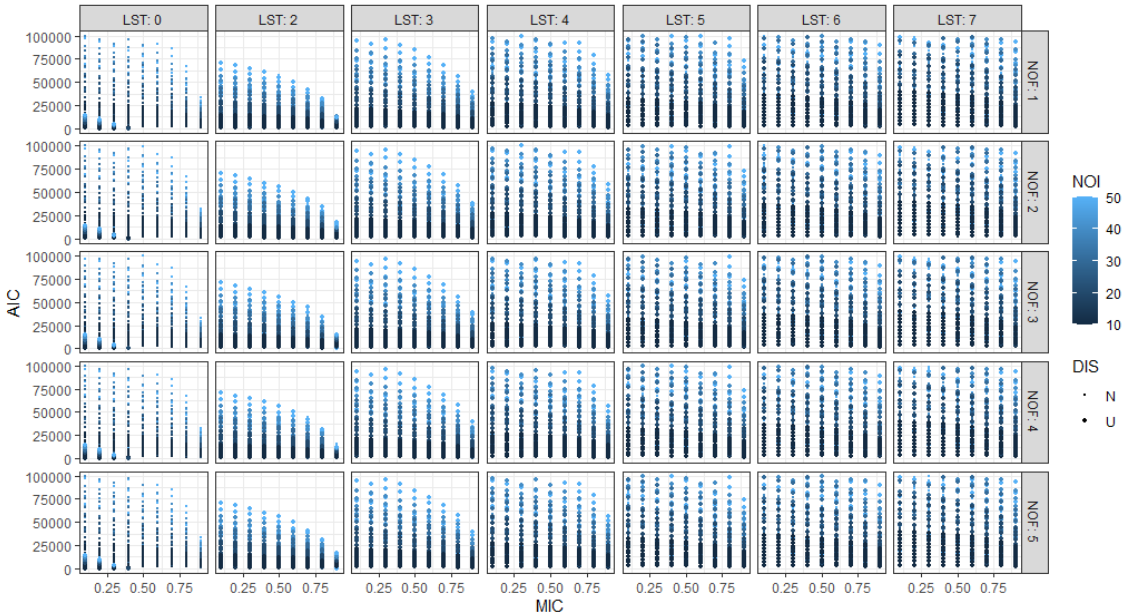
Rysunek 17. Poziom surowej wartości różnicy chi-kwadrat (χ^2_{diff}) dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST) oraz różnej liczby czynników (NOF) i liczby itemów (NOI)



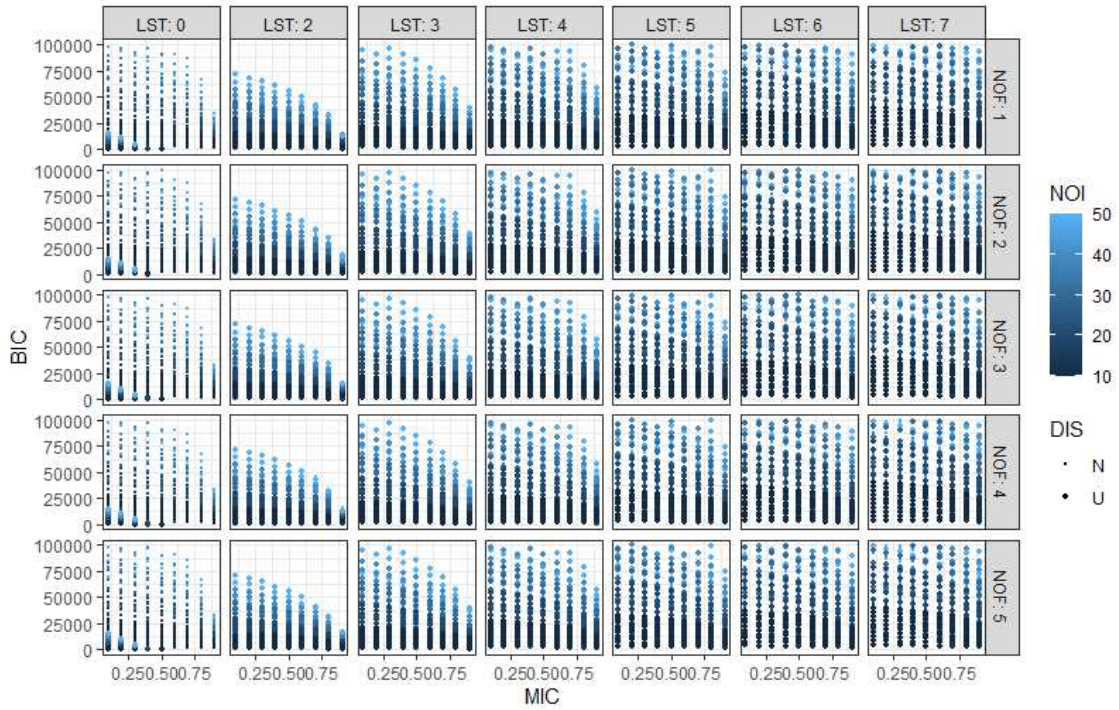
Rysunek 18. Poziom istotności dla różnicy chi-kwadrat ($\chi^2_{diff\text{-}p}$) dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST) oraz różnej liczby czynników (NOF) i liczby itemów (NOI)

Ostatecznie miary AIC i BIC, związane właściwie liniowo, dawały lepsze rezultaty w przypadku silniej skorelowanych itemów w obrębie skali i różniły się charakterystyką z racji rzutowania wartości rzeczywistych na skale Likerta, ale pomiędzy poszczególnymi modelami – jedno- i wielowymiarowymi – nie odnotowano różnic pozwalających na rozróżnienie, który model jest, a który nie jest prawidłowy (Rysunek 19, Rysunek 20).

Rysunek 19. Poziom AIC dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST) oraz różnej liczby czynników (NOF) i liczby itemów (NOI)



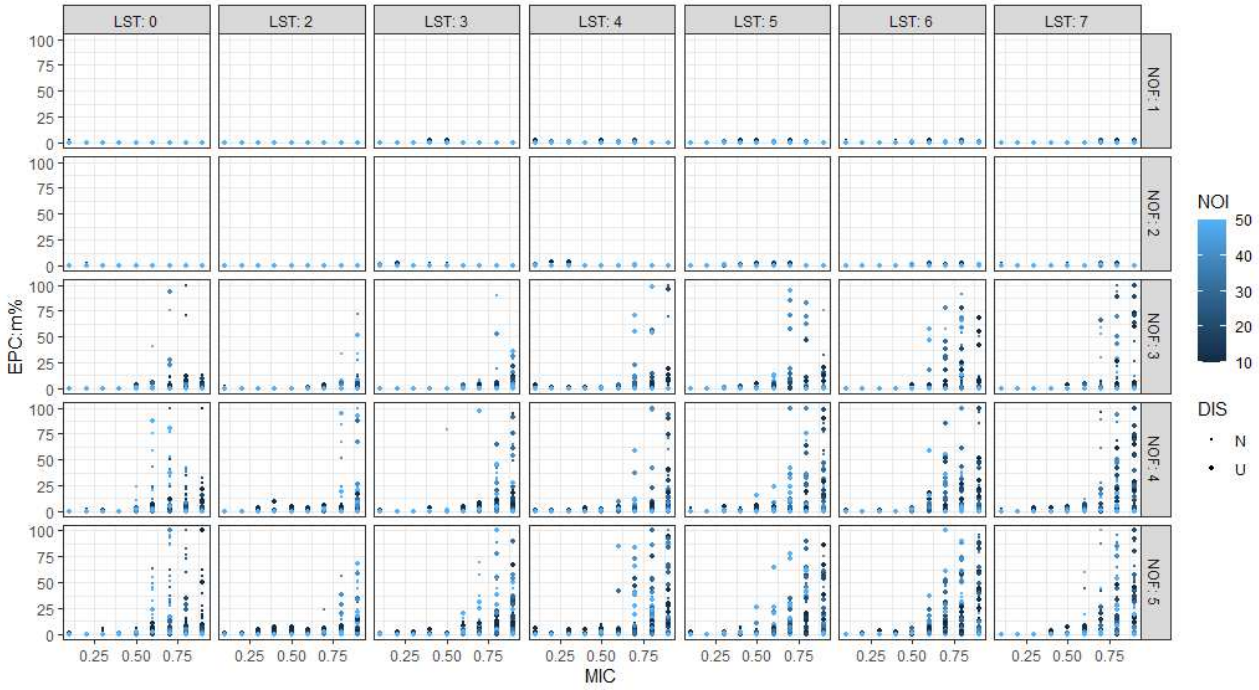
Rysunek 20. Poziom BIC dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST) oraz różnej liczby czynników (NOF) i liczby itemów (NOI)



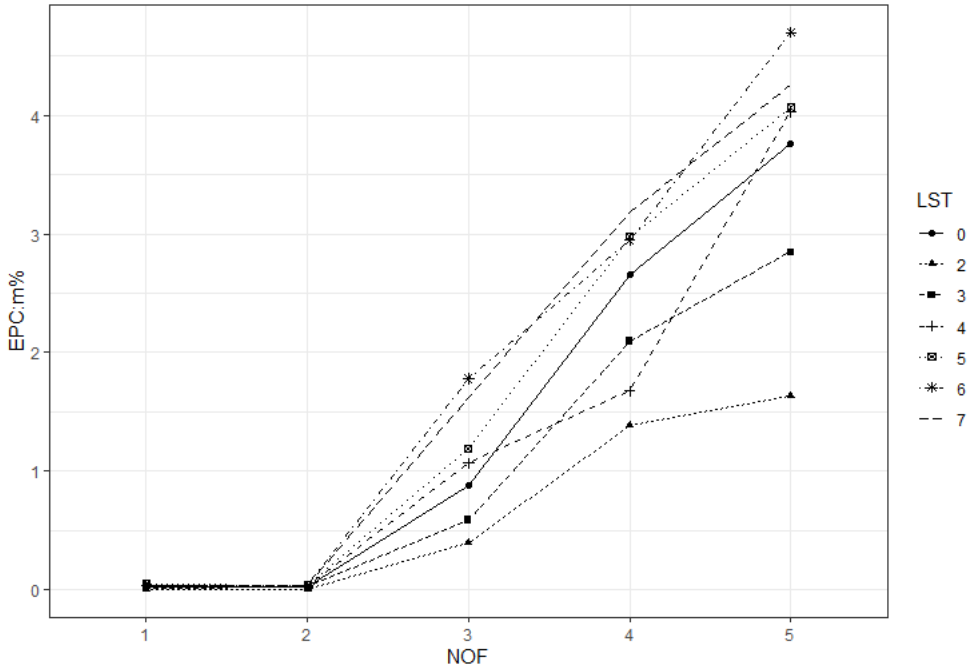
3.2.3.1. Indeksy modyfikacji

Popularne miary dopasowania, okazały się zatem nieszczególnie przydatne w określeniu poprawnej specyfikacji wymiarowości modelu. Pozostały do omówienia miary oparte o indeksy modyfikacji. Parametry *m%*, *epc:m%* i *total m%* zostały precyzyjniej przedstawione na wykresach poniżej. Jak widać (Rysunek 21) parametr *epc:m%* wprawdzie wzrastał nieznacznie wraz z liczbą (błędnych) dodatkowych wymiarów, ale nie pozwalał na prawidłowe rozróżnienie modeli jedno- i dwuwymiarowych (Rysunek 22), co wyklucza jego praktyczne zastosowanie. Silnie związane ze sobą parametry *m%* (Rysunek 23) i *total m%* (Rysunek 24) (rs = 0,930, $p < 0{,}001$) wzrastały wyraźnie wraz z wielkością struktury (liczbą czynników latentnych), ale te wzrosty były różne dla różnych typów skal (różnych długości skal Likerta). Ponadto, między modelami jedno- a dwuwymiarowymi przyrosty były bardzo małe w odniesieniu do pozostałych modeli (Rysunek 25), co ponownie oznacza, że rozróżnienie modeli jedno- i dwuwymiarowych nie jest możliwe.

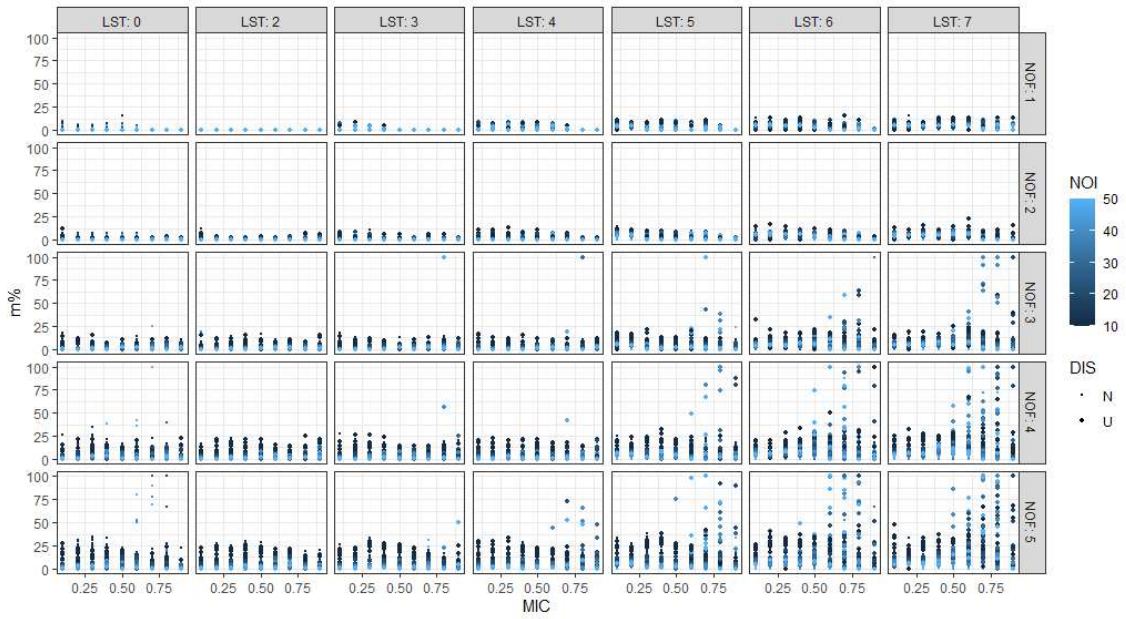
Rysunek 21. Wartości EPC:m% dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST) oraz różnej liczby czynników (NOF) i liczby itemów (NOI)



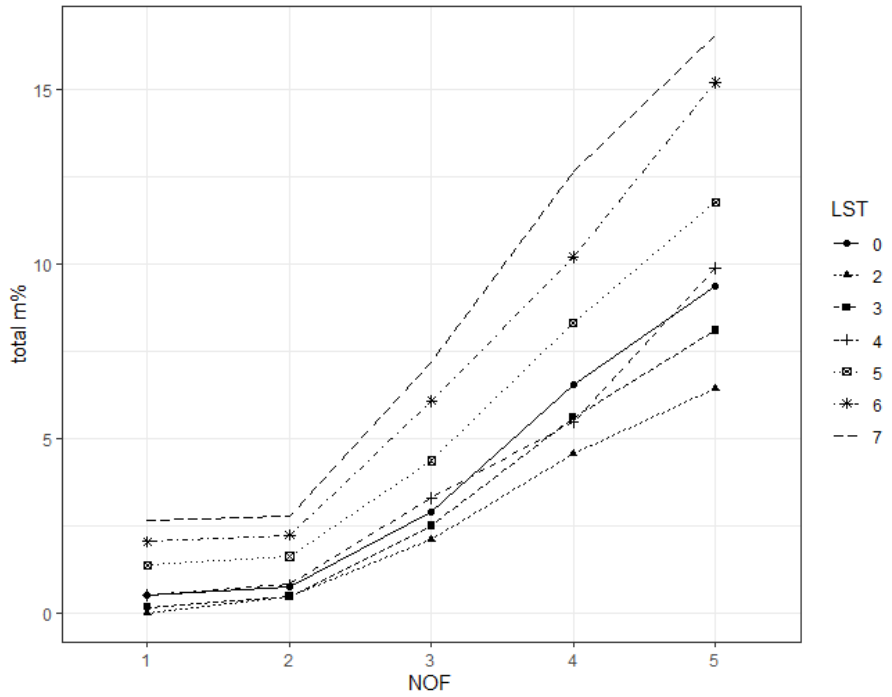
Rysunek 22. Wartości EPC:m% w zależności od różnych typów skal (LST) oraz różnej liczby czynników (NOF)



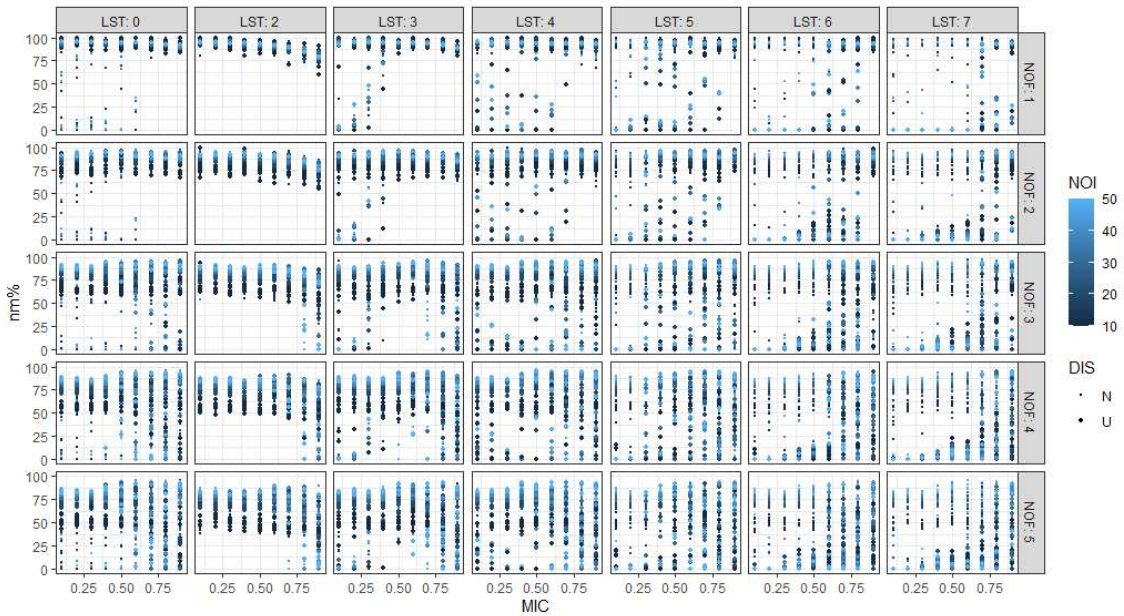
Rysunek 23. Wartości m% dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST) oraz różnej liczby czynników (NOF) i liczby itemów (NOI)



Rysunek 24. Wartości total m% w zależności od różnych typów skal (LST) oraz różnej liczby czynników (NOF)

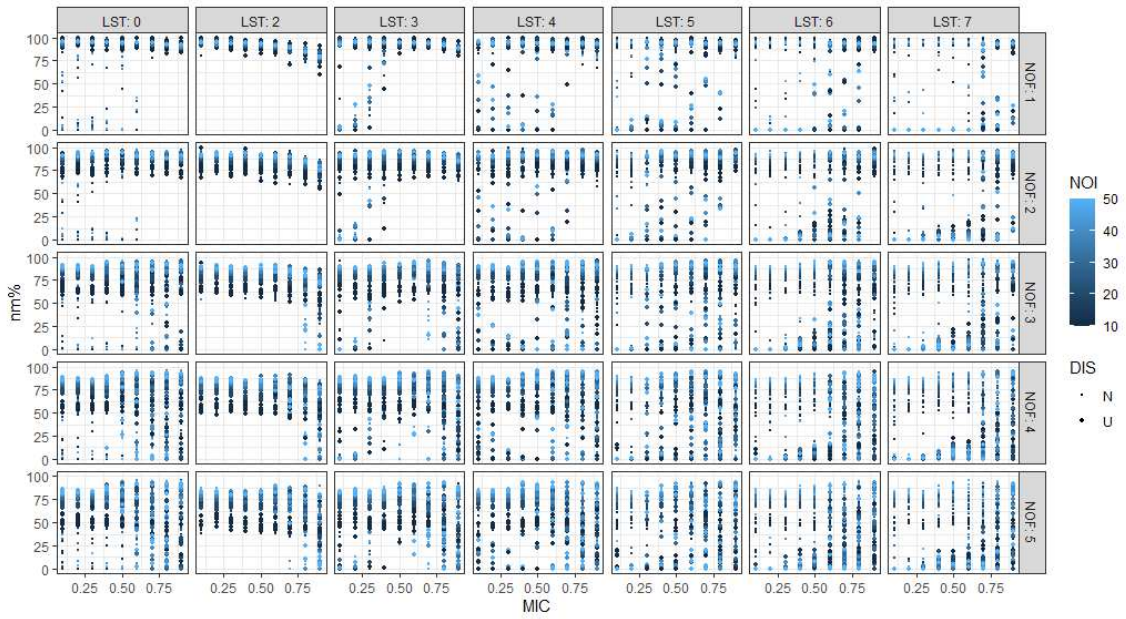


Rysunek 25. Wartości total.m% dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST) oraz różnej liczby czynników (NOF) i liczby itemów (NOI)

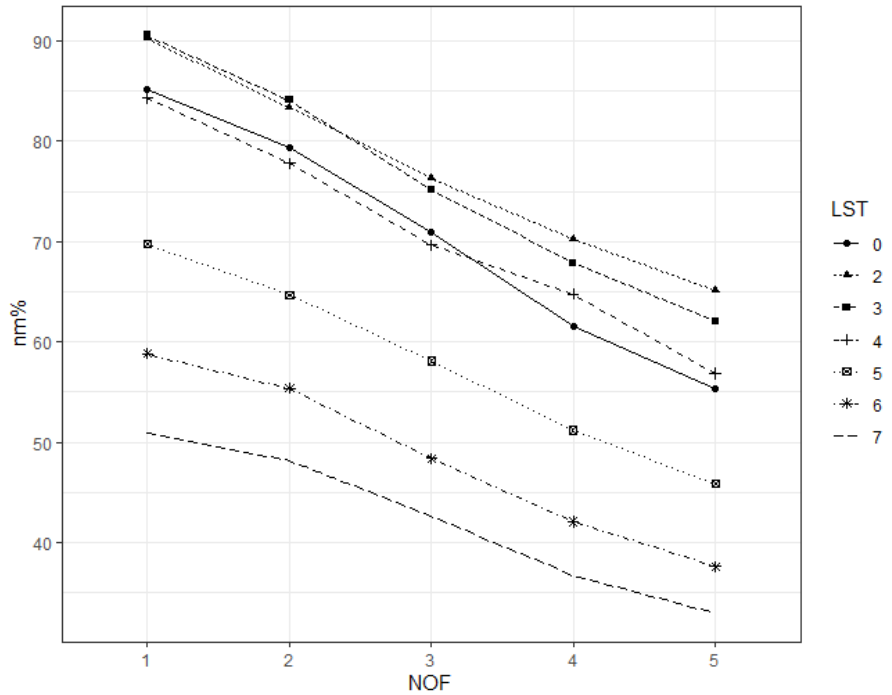


Średnie wartości kolejnych miar, czyli *nm%* (Rysunek 26, Rysunek 27), *ecp:nm%* (Rysunek 28, Rysunek 29) oraz *i%* (Rysunek 30, Rysunek 31), różniły się istotnie statystycznie pomiędzy modelami jednowymiarowymi a ogólnie wielowymiarowymi, ale dla *epc:nm%* modele jedno- i dwuwymiarowe ponownie były nierozróżnialne. Wartości średnie parametru *i%* nie mają praktycznego zastosowania, ponieważ dla dłuższych skal Likerta nie różnicowały modeli w ogóle, a modele jedno- i wielowymiarowe miały taki sam poziom niekonkluzywności co do potencjalnych ścieżek w modelu. Wraz ze wzrostem liczby czynników wyraźnie spadały wartości *nm%*, ale były to spadki relatywne, tzn. spadki w obrębie poszczególnych typów skal Likerta. Podsumowując można powiedzieć, że wartości *nm%* mogą posłużyć do porównywania ze sobą modeli w obrębie tego samego zestawu pytań kwestionariusza i skal Likerta o tych samych długościach, ale nie można tu wyznaczyć uniwersalnego punktu odcięcia bez uwzględnienia typu skali (Rysunek 27).

Rysunek 26. Wartości nm% dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST) oraz różnej liczby czynników (NOF) i liczby itemów (NOI)

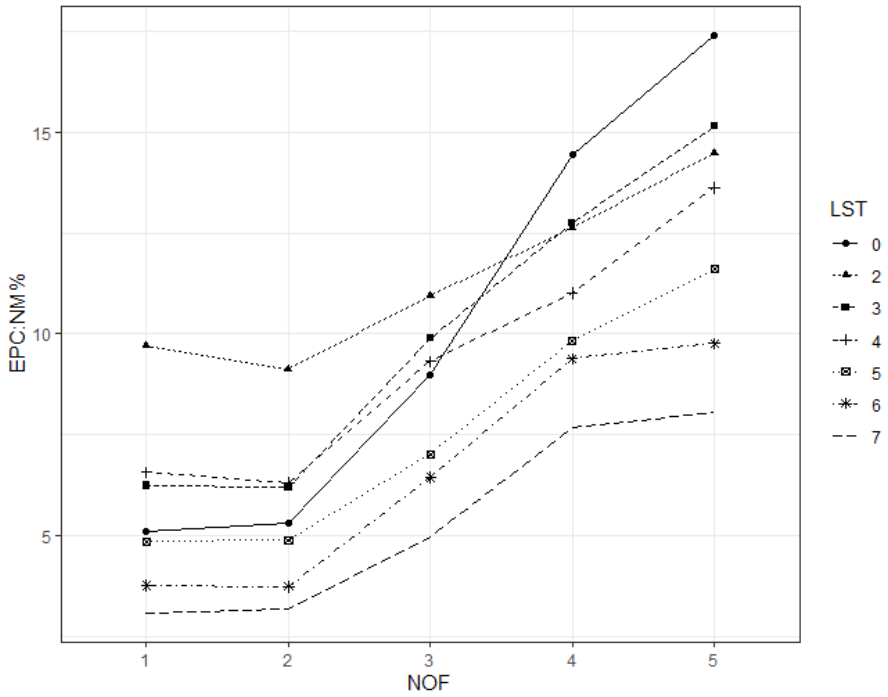


Rysunek 27. Wartości nm% w zależności od różnych typów skal (LST) oraz różnej liczby czynników (NOF)

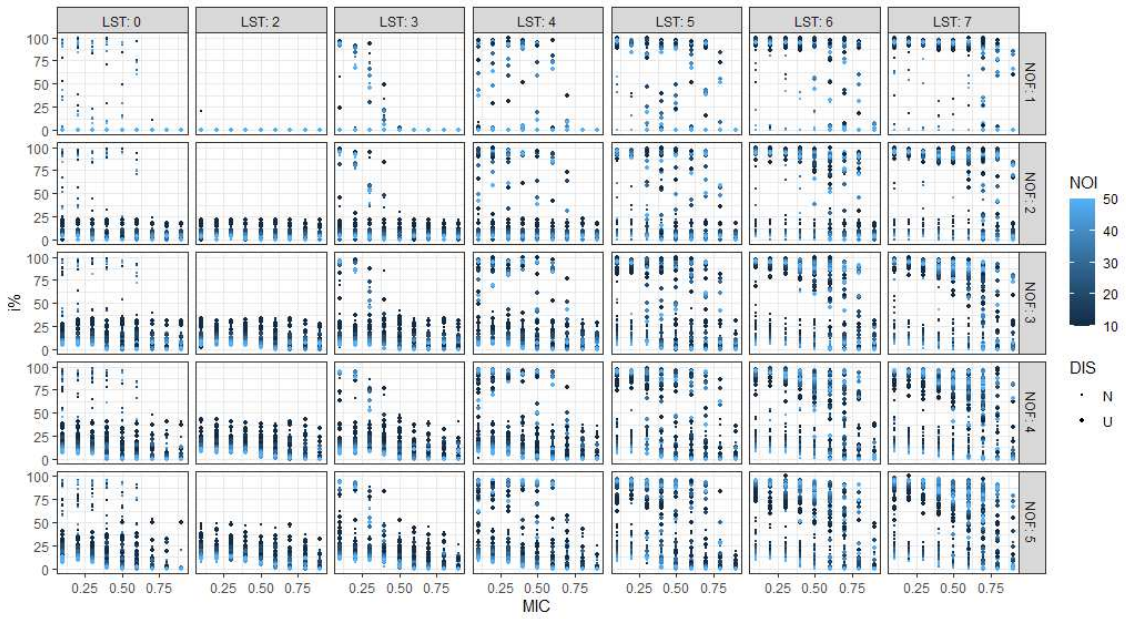


Rysunek 28. Wartości EPC:nm% dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST) oraz różnej liczby czynników (NOF) i liczby itemów (NOI)

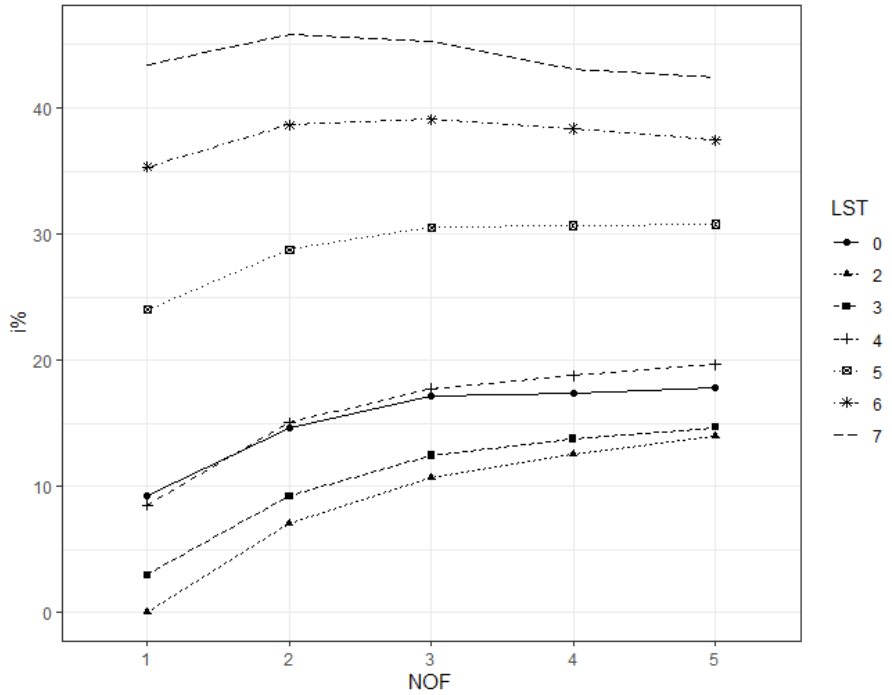
Rysunek 29. Wartości EPC:nm% w zależności od różnych typów skal (LST) oraz różnej liczby czynników (NOF)



Rysunek 30. Wartości i% dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST) oraz różnej liczby czynników (NOF) i liczby itemów (NOI)

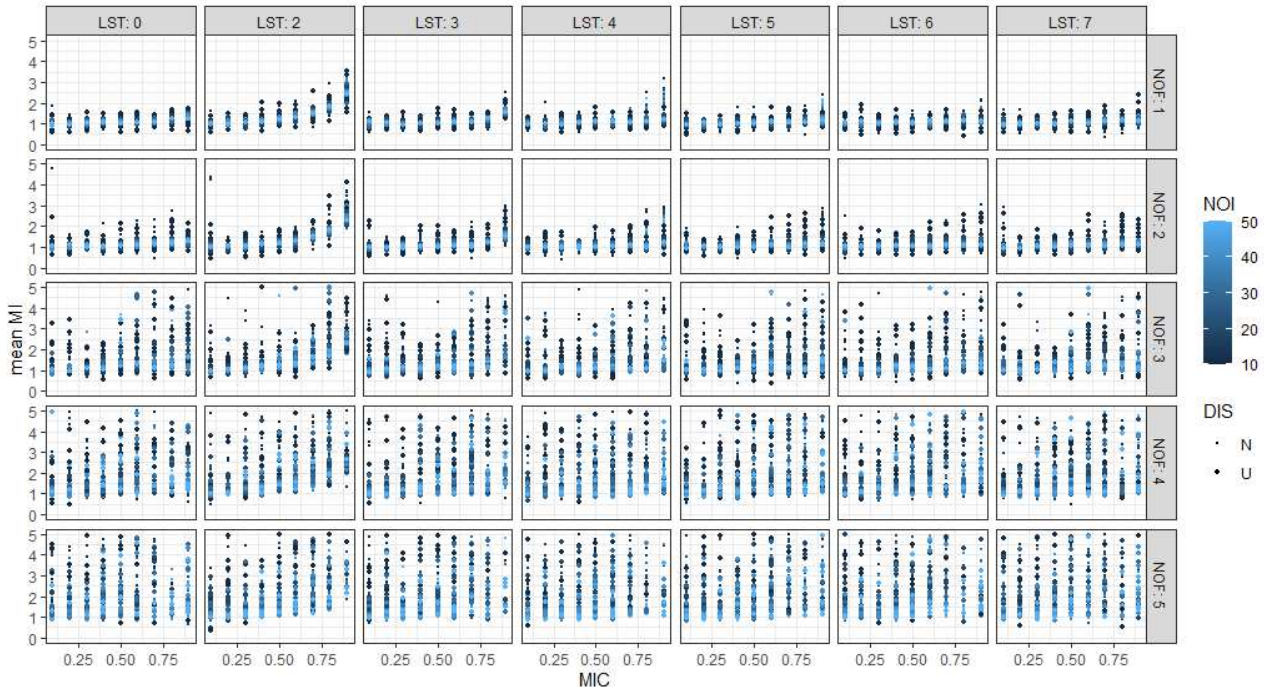


Rysunek 31. Wartości i% w zależności od różnych typów skal (LST) oraz różnej liczby czynników (NOF)

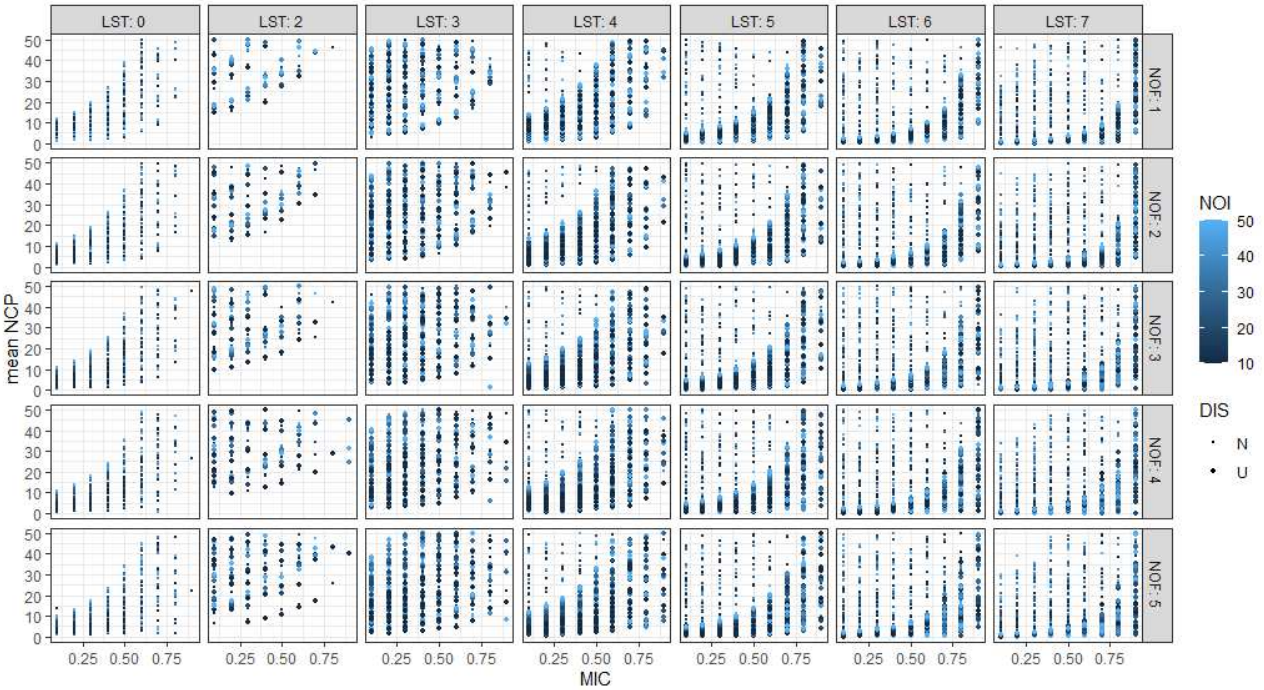


Na postawie średniej wartości *MI* (Rysunek 32) nie można rozróżnić poprawności modeli jedno- i dwuwymiarowych. Z kolei dla wartości średnich *NCP* (Rysunek 33) nie można określić poprawności pomiędzy modelami w obrębie jednego typu skali; na przykład tam gdzie były pięciostopniowe skale Likerta nie występowały różnice pomiędzy modelem jedno- a wielowymiarowym. Dla średnich wartości ECP (Rysunek 34) wartości powyżej 0,010 wskazywały na potencjalną nieprawidłową specyfikację modelu, niezależnie od typu skali. Jest to zatem jedyny wskaźnik, który daje wyraźny punkt odcięcia możliwy do praktycznego wykorzystania.

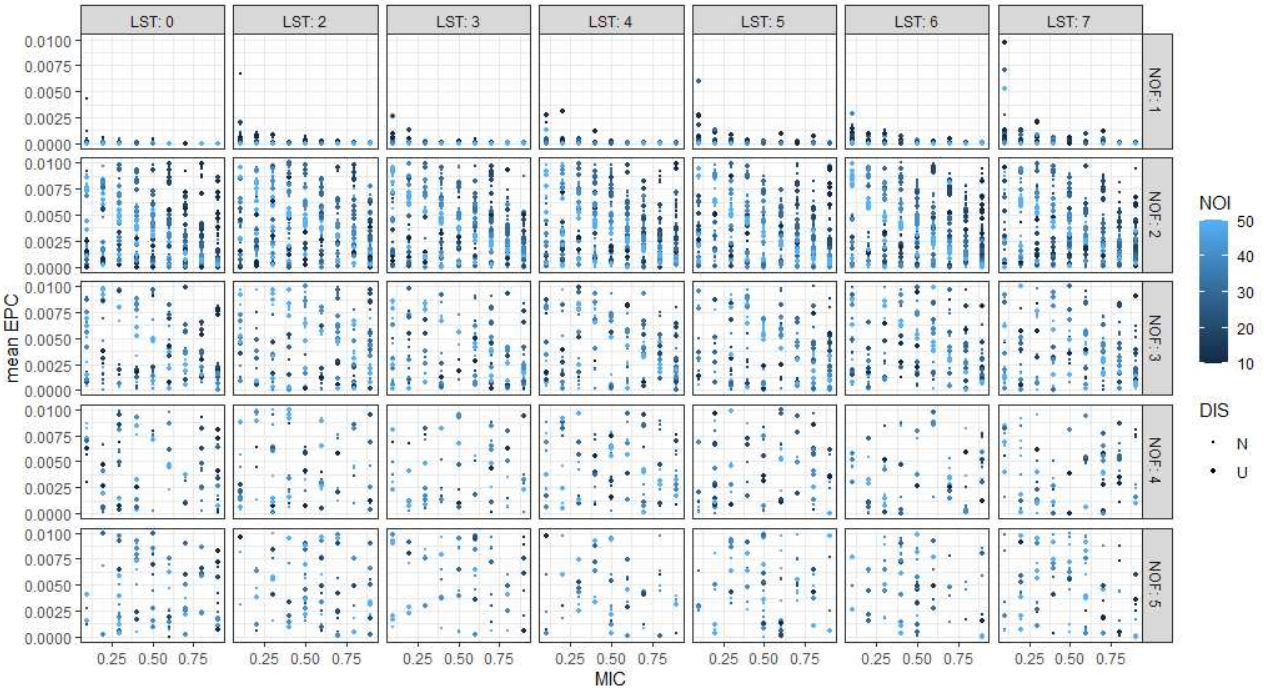
Rysunek 32. Poziom MI dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST) oraz różnej liczby czynników (NOF) i liczby itemów (NOI)



Rysunek 33. Poziom NCP dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST) oraz różnej liczby czynników (NOF) i liczby itemów (NOI)



Rysunek 34. Poziom EPC dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST) oraz różnej liczby czynników (NOF) i liczby itemów (NOI)



Podsumowując, właściwie wszystkie klasyczne miary dopasowania okazały się nieprzydatne w określeniu, czy poprawny jest model jedno- czy wielowymiarowy, co jest odpowiedzią na pytanie badawcze P2. Oczywiście nie chodzi tu jedynie o ustalenie z użyciem tych współczynników, czy model jest dopasowany do danych czy nie – po prostu w przypadku danych o niewątpliwie jednowymiarowej strukturze badacz potwierdzający swój model teoretyczny z tym samym prawdopodobieństwem może przyjąć za prawidłową strukturę niepoprawną (wielowymiarową) co poprawną (jednowymiarową). Interesujące wydają się być wyniki oparte o indeksy modyfikacji, z czego *nm%* można uznać za miarę pozwalającą na porównanie modeli w obrębie tej samej skali Likerta, a przeciętny poziom *EPC* większy niż 0,010 można uznać za wskazujący potencjalną nieprawidłową specyfikację modelu.

4. BADANIE 2 (DANE EMPIRYCZNE)

4.1. Metoda badania empirycznego

W celu replikacji uzyskanych wyników symulacyjnych na rzeczywistych danych, a tym samym zwiększenia pewności co do otrzymanych rezultatów, przeprowadzono badanie empiryczne z wykorzystaniem kwestionariuszy samoopisowych. Kluczowe było tu powtórzenie konkluzji dotyczących pytań badawczych P2 i P3. Dane zostały zgromadzone na obszernej próbie respondentów[12], co umożliwiło włączenie rzeczywistych obserwacji do algorytmu porównującego potencjalne struktury jedno- i wielowymiarowe w ramach analizy SEM. Podjęte działania miały na celu przeprowadzenie analogicznych analiz do tych, które wykonano dla modeli opartych na symulacjach, lecz tym razem bazując na rzeczywistych odpowiedziach uczestników badania, co pozwalało na dodatkową ocenę trafności uzyskanych wcześniej wyników.

Pewnego kłopotu nastręczał początkowo wybór narzędzi o niewątpliwie jednowymiarowej strukturze. Niewiele narzędzi psychometrycznych ma udokumentowany jeden silny czynnik. Ostatecznie wybrano kwestionariusz samooceny SES (Łaguna i in., 2007), kwestionariusz samoskuteczności GSES (Juczyński, 2000), oraz skalę do pomiaru lęku jako cechy kwestionariusza STAI (Sosnowski i Wrześniewski, 1983), ponieważ a) mają tę samą rozpiętość skal (wszystkie mają czterostopniowy system odpowiedzi na skali Likerta), b) zbliżoną liczbę itemów do warunków liczby itemów w badaniu symulacyjnym (GSES i SES mają po 10 itemów, STAI w skali lęku-cechy 20 itemów), oraz c) w badaniach empirycznych wykorzystywane są jako jednowymiarowe (na przykład u: Chen i in., 2004; Gazdowska, i in., 2015; Lightsey i in., 2006; Mur i in., 2022; Onyeizugbo, 2010; Jahanian i Poornaghi, 2012; Stanley i Murphy, 1997). Należy zwrócić uwagę na dwie kwestie: 1) ponieważ STAI ma dwa wymiary po 20 itemów, lęk-stan i lęk-cecha, wykorzystano tylko pierwszy z nich dla uproszczenia badań empirycznych; tu i we wszystkich dalszych częściach odniesienia do STAI oznacza tylko i wyłącznie lęk-stan; 2) wymieniona wcześniej literatura do kwestionariusza SES wskazuje na możliwość dwuczynnikowej struktury, lub innego zanieczyszczenia pojedynczego czynnika.

W analizach wykorzystano dane od N = 5172 osób w wieku od 18 do 80 lat (M = 30,23; SD = 10,49). Około dwie trzecie próby było kobietami (n = 3302; 66,1%), a jedna trzecia mężczyznami (n = 1697; 33,9%). Dwie trzecie osób badanych zadeklarowało

[12] Około 70% danych pochodziło z różnych moich wcześniejszych niepublikowanych badań

wykształcenie wyższe (*n* = 1996; 58,6%), około czterdziestu procent badanych średnie (*n* = 1331; 39,1%), mniej niż jedna dziesiąta uczestników badania miało wykształcenie podstawowe (*n* = 80; 2,3%). Kwestionariusz SES wypełniło 2640 osób, kwestionariusz GSES – 1503 osoby, a kwestionariusz STAI – 1029 osoby.

Badani, rekrutowani poprzez kody QR, linki oraz rekrutację bezpośrednią, byli zachęcani do wypełniania jednego z trzech kwestionariuszy (SES, STAI, GSES); każda osoba badana mogła wypełnić 1, 2 lub wszystkie 3 kwestionariusze, ale nie kontrolowano identyfikatorów między kwestionariuszami (nie było to badanie korelacyjne, a jedynie analiza struktur). W Tabeli 5 oraz Tabeli 6 zaprezentowano statystyki opisowe próby z badania empirycznego.

Tabela 5
Statystyki opisowe próby badania empirycznego

		Ogółem		STAI		SES		GSES	
		N	%	N	%	N	%	N	%
Płeć	Kobieta	3302	66,1%	1050	21,0%	1704	34,1%	548	11,0%
	Mężczyzna	1697	33,9%	450	9,0%	917	18,3%	330	6,6%
Wykształcenie	Podstawowe	80	2,3%	36	1,1%	12	0,4%	32	0,9%
	Średnie	1331	39,1%	429	12,6%	644	18,9%	258	7,6%
	Wyższe	1996	58,6%	886	26,0%	755	22,2%	355	10,4%

* $p < 0,05$;** $p < 0,01$; nie uzyskano od wszystkich badanych informacji demograficznych, stąd sumy poszczególnych podgrup są zwykle mniejsze niż $N = 5172$

Tabela 6
Statystyki opisowe wieku badanych w badaniu empirycznym (N = 5172)

	Wiek	M	SD	Mdn	Sk	Kurt	D
Ogółem	18-80	30,23	10,49	26	1,43	1,73	0,18**
Kobieta	18-80	29,53	10,01	26	1,54	2,13	0,19**
Mężczyzna	18-80	31,74	11,33	28	1,22	1,04	0,16**
STAI	18-63	28,56	9,61	27	1,01	0,67	0,12**
GSES	18-77	34,46	11,93	31	1,02	0,44	0,14**
SES	18-80	28,68	9,39	25	1,80	3,05	0,22**

* $p < 0,05$;** $p < 0,01$

Dążono do wykonania analiz analogicznych do badania symulacyjnego. Niemniej należy zauważyć, że niektóre parametry badania symulacyjnego w warunkach naturalnych były pomijane z racji struktury samych kwestionariuszy. Zestawione w Tabela 1 warunki symulacji

musiały zostać ograniczone następująco: a) rodzaj skal Likerta i typ rozkładu były podyktowane realnymi danymi, to znaczy rozpiętość skali była oryginalną rozpiętością skali dla źródłowej wersji badanego kwestionariusza – wszystkie cztery kwestionariusze miały czterostopniowy system odpowiedzi na skali Likerta; b) założona odgórnie korelacja w symulacji stała się w badaniu po prostu realnym poziomem średnich korelacji w uzyskanym podzbiorze danych kwestionariuszowych; c) liczba itemów kwestionariusza była stała i zgodna z oryginalną liczbą tych itemów – SES i GSES miały ich po 10, STAI miał ich 20.

Następnie, po zebraniu i opracowaniu danych, trzy zbiory danych stały się podstawą algorytmu podobnego do metody Monte Carlo, z tą różnicą, że zamiast symulować pod-zbiory danych, pobierano po prostu losową próbę z danych realnych. Na przykład, z zawierającego 2640 obserwacji zbioru danych z wynikami kwestionariusza SES losowano podzbiór o zadanej liczbie obserwacji zgodnej z warunkami badawczymi z Tabela 1: od 100 do 1000 obserwacji. Podobnie jak w badaniu symulacyjnym, ustalano losowo liczbę testowanych czynników latentnych (od 1 do 5). Dla danych pobranych z badania empirycznego było więc 50 zgodnych z Tabela 7 możliwych warunków badawczych (10 wielkości próby x 5 testowanych struktur). Dla każdego z warunków badawczych wylosowano po 100 zbiorów danych, co dało przeprowadzonych 5000 symulacji na każde kwestionariuszowe narzędzie badawcze (razem 15000 punktów danych).

Tabela 7
Warunki wejściowe badania empirycznego

Warunek (czynnik)	Poziomy zmiennej	skrót
Rozkład	--- pominięty i zgodny z wynikami empirycznymi ---	DIS
Średnia korelacja między pozycjami	--- pominięty i zgodny z wynikami empirycznymi ---	MIC
Liczba itemów kwestionariusza	10 (GSES, SES), 20 (STAI)	NOI
Wielkość próby	100, 200, 300, 400, 500, 600, 700, 800, 900, 1000	SMP
Typ skal Likerta	--- zgodnie z oryginalnym narzędziem, 4 ---	LST
Liczba zmiennych latentnych	1, 2, 3, 4, 5	NOF
Ocena struktury	UNIDIM, NOT-UNID	AOS

Adnotacja: „skrót” oznaczają skróty użyte na wykresach i w tabelach; UNIDIM oznacza jednowymiarową strukturę; NOT-UNID oznacza niejednowymiarową strukturę

Dalszy przebieg generowania wyników był zgodny z generowaniem wyników dla badania symulacyjnego: losowano ze zbioru danych empirycznych podzbiór o zadanej wielkości; wybierano, czy testowana będzie struktura jedno- czy wielowymiarowa (od 2 do 5 wymiarów); a następnie obliczano wszystkie współczynniki dopasowania modeli SEM i indeksy modyfikacji. Wynik takich obliczeń tworzył punkt danych, a 5000 punktów danych

dla każdego z trzech kwestionariuszy tworzyło macierz wyników (razem trzy macierze wyników dla trzech narzędzi badawczych dające 15000 punktów danych).

W pierwszej kolejności oceniono jednowymiarowość uzyskanych podzbiorów z danymi – użyto tych samych narzędzi oceny co dla zbiorów danych symulowanych. W pierwszej kolejności dokonano zatem sprawdzenia poziomu jednowymiarowości analizowanych wyników (tabela poniżej).

Tabela 8
Procent rozpoznania modeli w danych empirycznych jako jednowymiarowych z użyciem różnych metod (N = 15000)

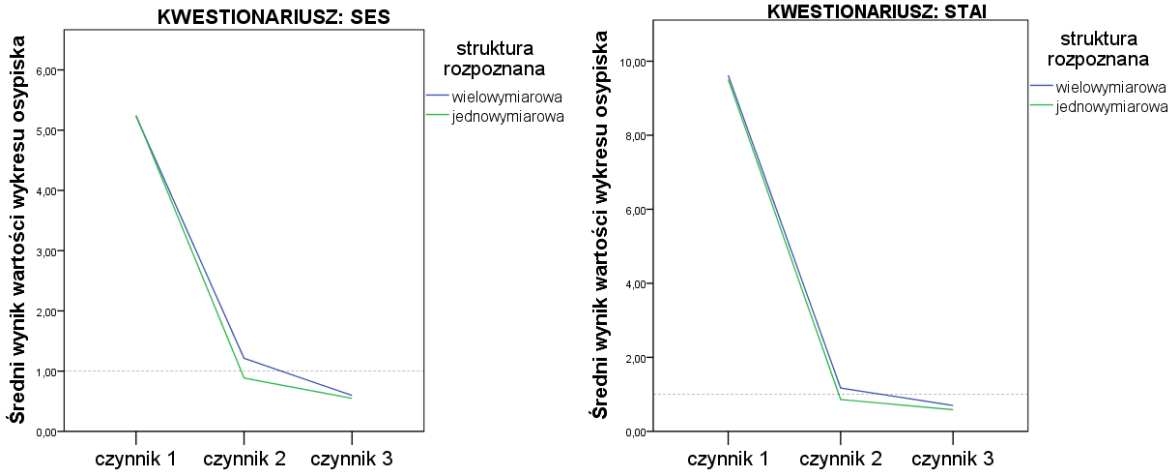
	GSES	SES	STAI	Ogółem
Kryterium Cattela	88,92%	63,96%	100,00%	84,29%
Analiza równoległa	100,00%	7,20%	17,48%	41,56%
Osypisko analizy równoległej	100,00%	99,80%	99,94%	99,91%
Test MAP	100,00%	5,80%	7,06%	37,62%
Sekwencyjny χ2	1,34%	0,00%	0,00%	0,43%

Okazało się, że realne dane mają niższy poziom pewności co do jednowymiarowości niż dane symulowane (co w zasadzie nie wydaje się szczególnie zaskakujące). Należy zaznaczyć, że odsetek rozpoznań tym razem bardzo się różnił pomiędzy poszczególnymi metodami użytymi do oceny oraz samymi testowanymi kwestionariuszami. Jeśli chodzi o strukturę skali GSES to w zasadzie wszystkie cztery metody oceny wymiarowości wskazały, że właściwie w każdym wypadku można ją traktować jako skalę jednowymiarową. Ponownie, sekwencyjny test χ2 nie dawał praktycznie żadnych sensownych rezultatów, należy go więc pominąć w ocenie (wielo)wymiarowości.

W przypadku kwestionariusza SES wykres osypiska i wykres analizy równoległej dały dość wysokie odsetki rozpoznania, ale sama analiza równoległa i test MAP wskazały niski procent rozpoznań struktury jako jednowymiarowej. Dla kwestionariusza STAI kryteria wizualne wskazywały praktycznie zawsze jeden czynnik, ale analiza równoległa i test MAP wskazywały raczej modele wielowymiarowe. Ponieważ wynik samej automatycznej analizy trudno sprawdzić pod kątem tego, czy jest to jeden zanieczyszczony czynnik, czy więcej niż jeden czynnik, potraktowano dalej niekonkluzywne modele jako po prostu niejednowymiarowe. Zakres tych rozbieżności wydaje się być interesującym zagadnieniem

przyszłych badań, zwłaszcza biorąc pod uwagę wykresy średnich wartości własnych pierwszych trzech czynników dla SES i STAI (Rysunek 35): różnica między strukturą jedno- a dwuczynnikową jest mała (co wskazywałoby raczej na zanieczyszczenia pojedynczego czynnika niż wielowymiarowość narzędzi badawczych). Niemniej, nie jest to przedmiotem badania niniejszej pracy. Do rozstrzygnięcia wybrano kryterium analizy równoległej, które, z racji swojej mocy niezależnej od wielości próby, powinno mieć najwyższe prawdopodobieństwo wskazania prawidłowo struktury jednowymiarowej, nawet jeśli jest ona zanieczyszczona (zob. Auerswald i Moshagen, 2019). Powstała w ten sposób dodatkowa zmienna grupująca na dwa wspomniane typy modeli uzyskanych z losowania w danych empirycznych.

Rysunek 35. Przeciętne (uśrednione) wykresy osypiska dla modeli uznanych za jedno- i wielowymiarowe w badaniach empirycznych



Adnotacja: Uśredniono wartości trzech pierwszych punktów (czynników) uzyskiwanych w analizie wykresu osypiska metodą Cattela; dalsze czynniki w obu przypadkach były pomijalne. Przerywana linia wskazuje standardowy punkt odcięcia (y = 1).

Podsumowując: zebrano dane empiryczne od osób badanych, przeprowadzono na nich analizy zbliżone do tych w symulacjach, a wyniki przedstawiono w podziale na modele jedno- i wielowymiarowe.

4.2. Wyniki badania empirycznego

Dla uzyskanych wyników z badania empirycznego można przedstawić trzy razy więcej analiz i wyników niż dla badania symulacyjnego ponieważ uzyskano wyniki dla trzech różnych kwestionariuszy badawczych. Postanowiono więc skupić się na kluczowych wnioskach, a dane prezentować na wizualizacjach zbiorczo dla wszystkich trzech narzędzi badawczych w celu skrócenia wywodu o powtarzalne wyniki. Wszystkie wyniki są oczywiście dostępne

w repozytorium niniejszej pracy pod linkiem https://osf.io/y36hz wraz z kodem w R pozwalającym na wygenerowanie wysokojakościowych wykresów.

Badanie empiryczne miało za zadanie potwierdzenie wyników szczegółowych badania symulacyjnego w zakresie przede wszystkim pytań badawczych P2 i P3, które miary dopasowania i oceny modelu CFA pozwalają na najbardziej jednoznaczne wnioski dotyczące akceptacji bądź odrzucenia modelu w przypadku, gdy konstrukt jest jednowymiarowy, ale badacz nie jest pewny ile ma wymiarów, oraz czy da się wskazać punkty odcięcia dla oceny, czy model jest rzeczywiście jednowymiarowy.

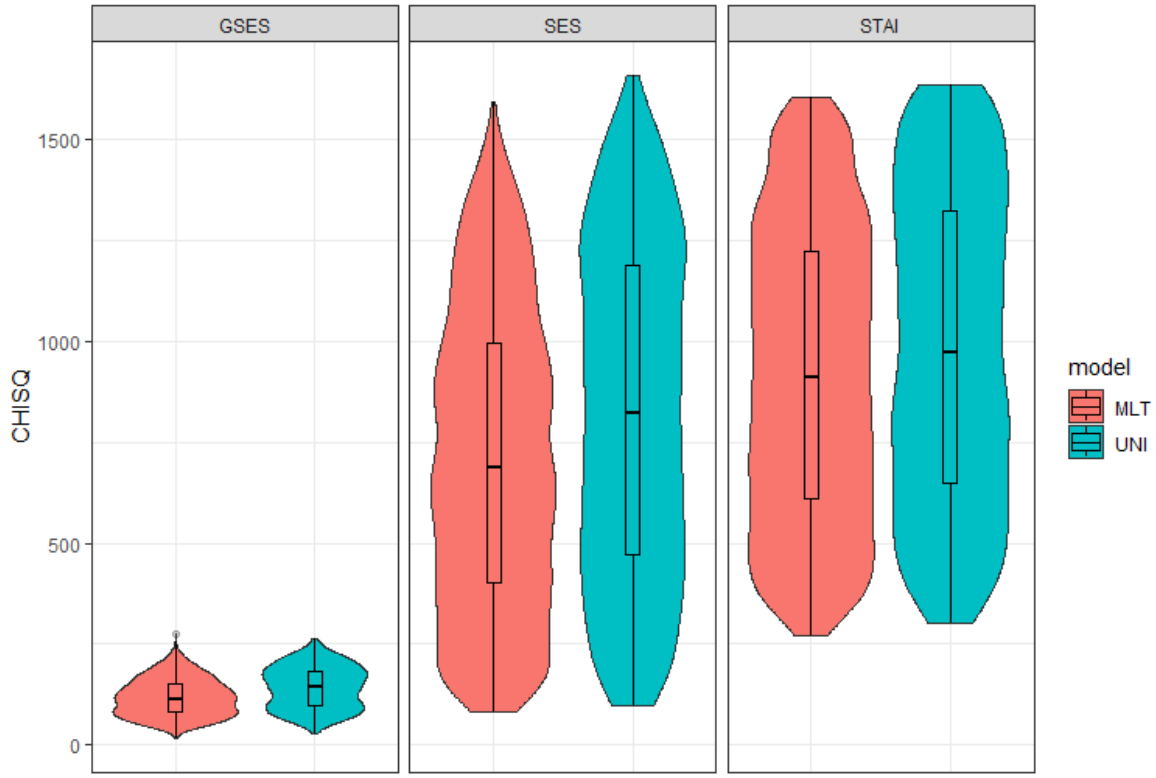
4.2.1. Użyteczność miar dopasowania w rozstrzyganiu o wymiarowości

4.2.1.1. Miary globalne

Najpierw zostaną zaprezentowane globalne miary dopasowania, a raczej różnic między wynikami uzyskanymi dla modeli jedno- i wielowymiarowych. Prezentacja wizualna poniżej została opracowana z użyciem wykresów wiolinowych, gdzie szerokość kolorowego obszaru wskazuje częstość występowania danego wyniku, wewnątrz znajdują się słupki średnich oraz błędu. Wykresy zostały podzielone na trzy okna – osobno dla każdego kwestionariusza badawczego, a kolorami oznaczono wartości dla modeli wielowymiarowych (kolor czerwony) i modeli jednowymiarowych (niebieski).

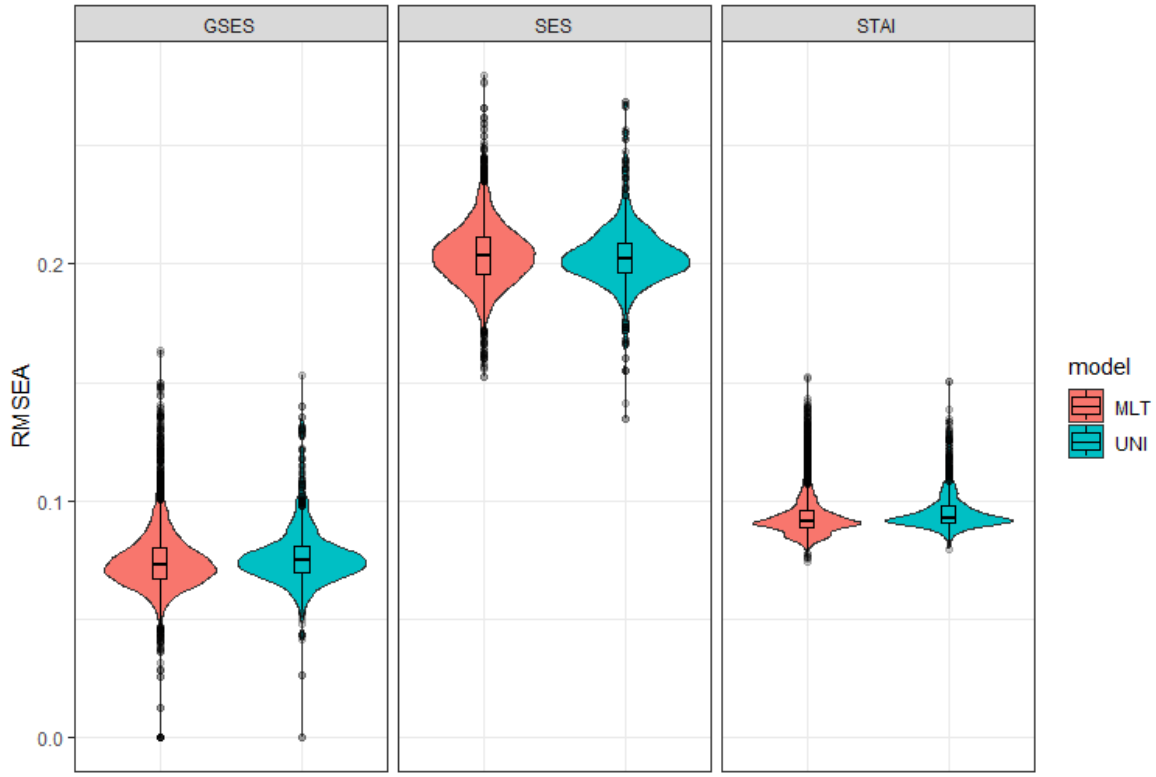
Biorąc pod uwagę zbiorczo dane dopasowywane jako jedno- vs wielowymiarowe dla wszystkich trzech kwestionariuszy wartość surowa $\chi 2$ (Rysunek 36) oraz wartość surowa RMSEA (Rysunek 37) były zbliżone i szansa dopasowania poprawnego modelu jako jedno- lub wielowymiarowego była właściwie taka sama. Podobnie zatem jak w badaniu symulacyjnym, jednoznaczne ustalenie, czy struktura jest jedno- czy wielowymiarowa z użyciem tych miar jest niemożliwa, bo można je tak samo dobrze (lub akceptowalnie) dopasować. Okazało się przy tym, że dane empiryczne z kwestionariusza SES mają ogólnie mały poziom dopasowania do modelu teoretycznego w porównaniu do kwestionariuszy GSES i STAI biorąc pod uwagę wartości RMSEA

Rysunek 36. Poziomy χ^2 dla modeli jedno- (UNI) i wielowymiarowych (MLT) uzyskanych z danych empirycznych



.

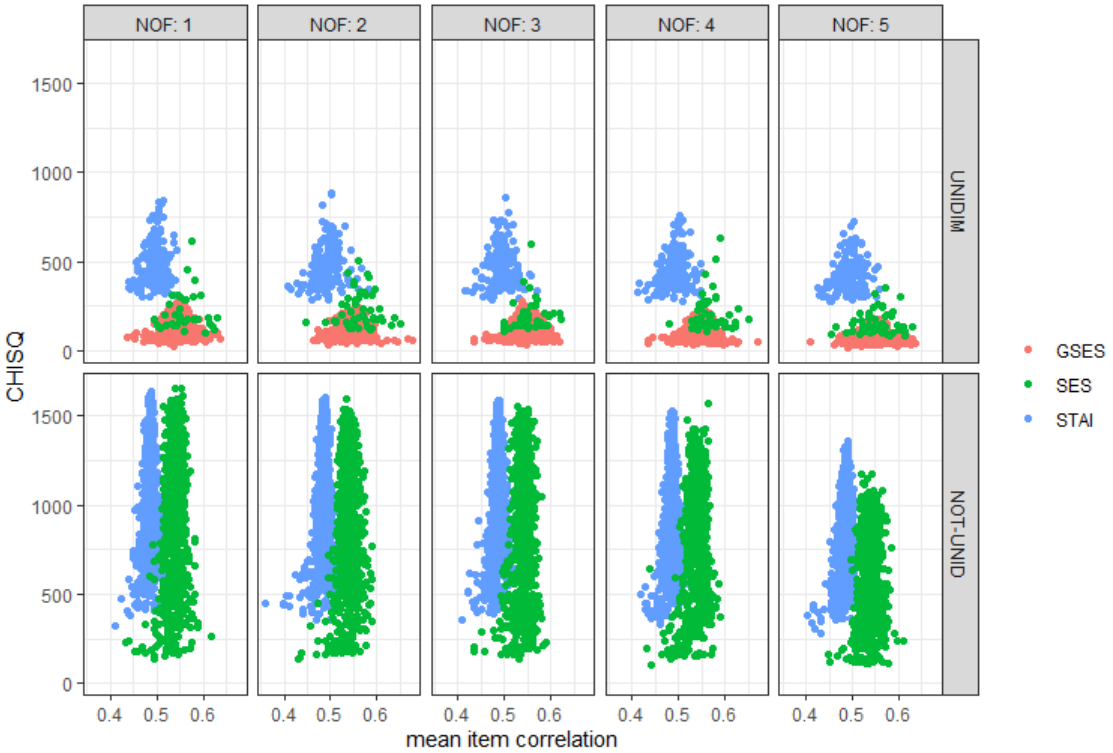
Rysunek 37. Poziomy RMSEA dla modeli jedno- (UNI) i wielowymiarowych (MLT) uzyskanych z danych empirycznych



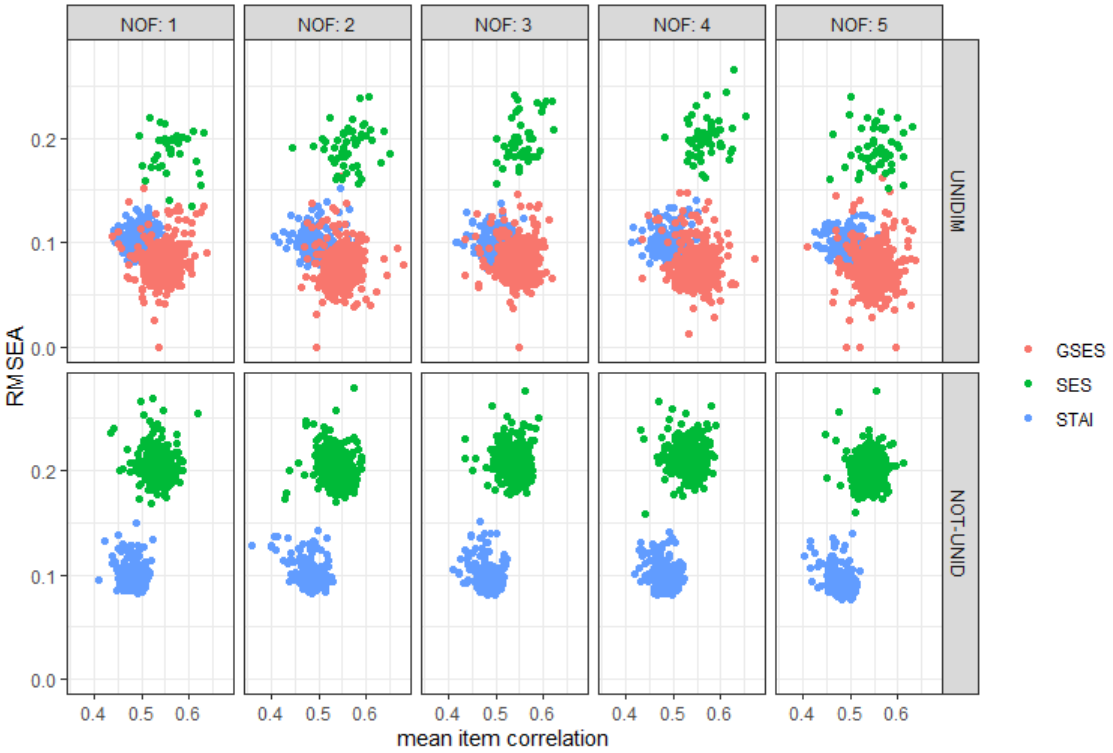
Analizując wyniki bardziej szczegółowo, można rozpatrzyć ich rozrzut, uwzględniając liczbę dopasowywanych czynników, średnią uzyskaną korelację w obrębie kwestionariusza (oś *X*) i jednocześnie rozpoznanie z użyciem wymienionych metod tego, czy struktura jest, czy nie jest jednowymiarowa. Zatem następny wykres i wszystkie analogiczne poniżej zostały podzielone na okna, dla których w wierszu górnym zaprezentowano wyniki rozpoznane jako jednowymiarowe, w wierszu dolnym zaprezentowano wyniki nierozpoznane jako jednowymiarowe (niekonkluzywne, zanieczyszczone lub wielowymiarowe); w kolejnych pięciu kolumnach znajdują się różne liczby dopasowywanych czynników (od jednego do pięciu), a kolorami zaznaczono wyniki dla różnych narzędzi badawczych. W osi *Y* wszystkich okien wykresów jest wartość analizowanego współczynnika dopasowania (na wykresie poniżej jest to miara χ^2). Na wykresie poniżej i na następnych analogicznych wykresach należy zauważyć, że wszystkie modele dla kwestionariusza GSES były z użyciem dowolnych metod rozpoznane jako jednowymiarowe, dlatego nie znajdują się w drugim wierszu wykresu. Ponadto, we wszystkich przypadkach miary dopasowania GSES były lepsze niż miary dopasowania pozostałych kwestionariuszy. Liczba itemów na wykresach nie jest uwzględniona, ponieważ liczbą itemów różnią się wykorzystane kwestionariusze. Jest to więc zmienna kategoryzująca o całkowicie redundantnym dla wyników charakterze

Na Rysunek 36 można zauważyć, że miara χ^2 nieznacznie się poprawia wraz liczbą czynników, przy czym dla modeli rozpoznanych jako jednowymiarowe ten spadek jest bardzo mały, a w przypadku modeli rozpoznanych jako niejednowymiarowe wraz ze wzrostem liczby czynników spadki są wyraźniejsze. Na Rysunek 39 widać, że wartości RMSEA właściwie się nie różnią od siebie we wszystkich typach modeli – ani rozpoznanie typu modelu, ani liczba czynników nie miały tu wpływu na całkowicie losowe rozrzuty dopasowań. Potwierdza to wcześniejsze obserwacje symulacyjne.

Rysunek 38. Przeciętne wartości χ^2 dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnej liczby czynników (NOF) i rozpoznania typu modelu (jedno- i niejednowymiarowe) (AOS)

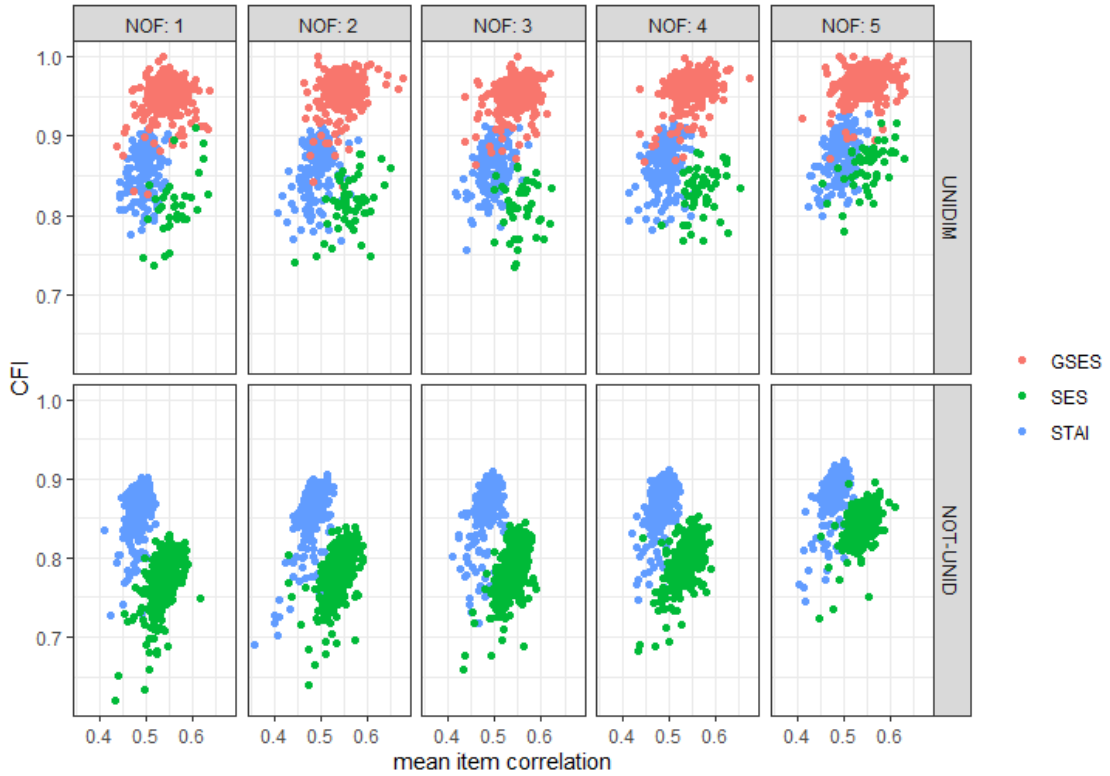


Rysunek 39. Przeciętne wartości RMSEA dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnej liczby czynników (NOF) i rozpoznania typu modelu (jedno- i niejednowymiarowe) z użyciem metod automatycznych (AOS)



Miary CFI, TLI, GFI, AGFI, oraz RNI były ze sobą silnie dodatnio skorelowane (Tabela 13), dlatego przedstawiony zostanie tu szczegółowo tylko poziom CFI (Rysunek 40); pozostałe wyniki były niemal identyczne dla wszystkich tych miar. W przypadku modeli rozpoznanych jako jednowymiarowe – niezależnie od liczby czynników poziom dopasowania był zbliżony, choć warto tutaj dodać, że kwestionariusz GSES miał wszędzie takie same, dość wysokie poziomy dopasowania, natomiast dla SES i STAI nieco częściej zdarzały się modele źle dopasowane. W przypadku modeli rozpoznanych jako niejednowymiarowe wraz z poziomem średniej korelacji pomiędzy pozycjami nieznacznie poprawiały się współczynniki dopasowania. Im większa była liczba czynników w modelach niejednowymiarowych, tym rzadziej zdarzały się poziomy miar świadczące o niedopasowaniu, a przyrost ten był nieco większy niż w przypadku modeli jednowymiarowych.

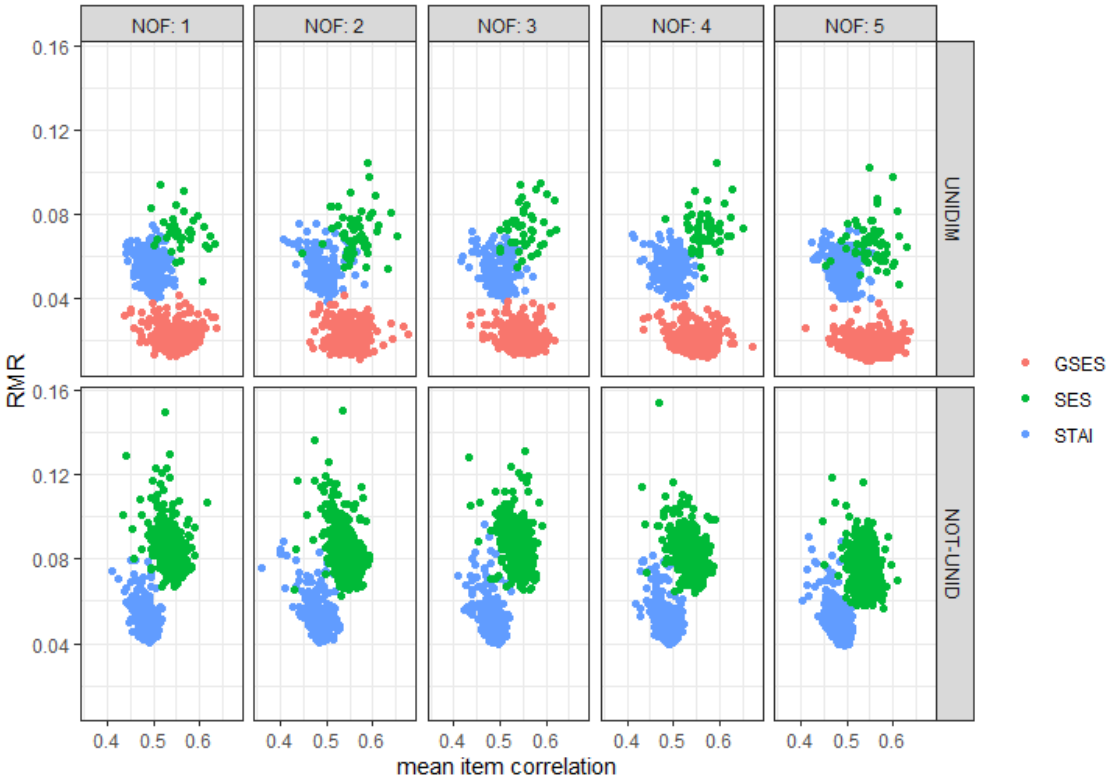
Rysunek 40. Przeciętne wartości CFI dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnej liczby czynników (NOF) i rozpoznania typu modelu (jedno i niejednowymiarowe) z użyciem metod automatycznych (AOS)



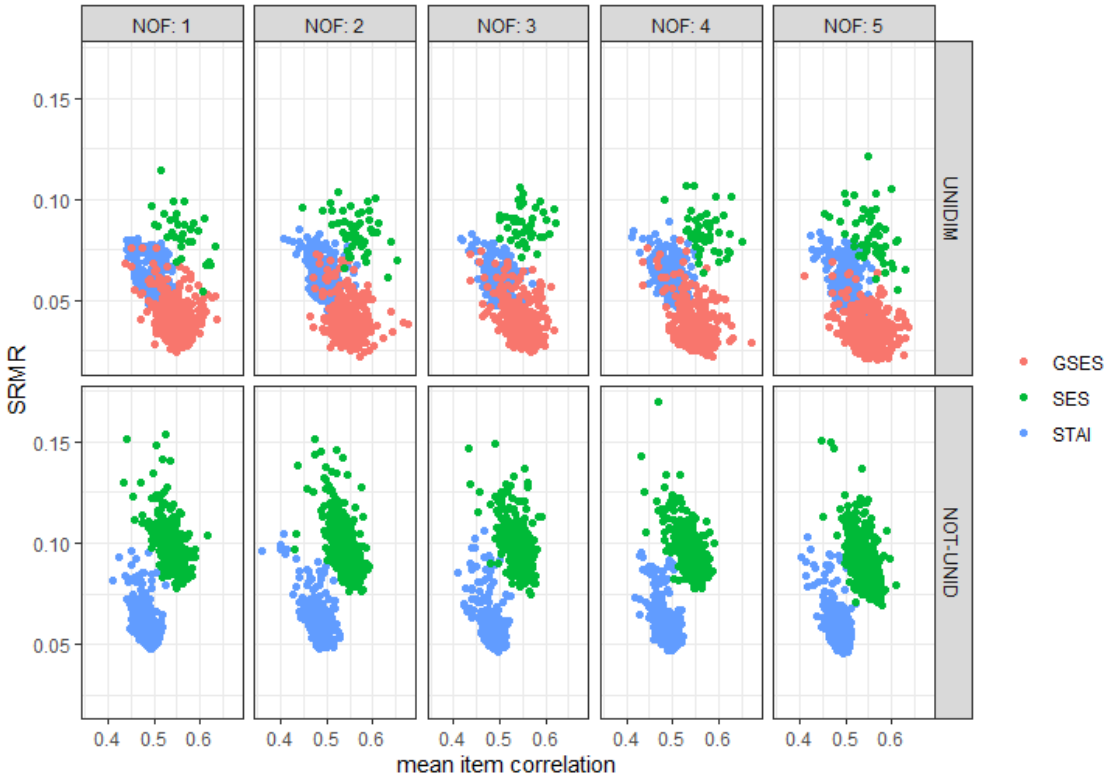
Miary RMR i SRMR nie korelowały silnie ze sobą, nie korelowały też z innymi współczynnikami, dlatego są przedstawione osobno na Rysunek 41 i Rysunek 42. W obu przypadkach w obrębie modeli rozpoznanych jako jednowymiarowe poziomy dopasowania były niemal takie same dla dowolnej liczby czynników i średniej korelacji między pozycjami, przy czym kwestionariusz GSES był dopasowany w każdym wariancie w lepszym stopniu niż

pozostałe kwestionariusze (które i tak można byłoby przy bardziej liberalnych wymaganiach uznać za akceptowalne). W przypadku struktur rozpoznanych jako niejednowymiarowych dopasowanie mierzone RMR i SRMR było ogólnie nieco słabsze oraz wzrastało ze średnim poziomem skorelowania itemów kwestionariusza. Nieznacznie też poprawiały się współczynniki wraz ze wzrostem liczby czynników, ale ta poprawa jest mała i w zasadzie rzadziej zdarzały się modele źle dopasowane, a liczba dopasowanych poprawnie (poniżej progów akceptowalności) właściwie się nie zmieniała.

Rysunek 41. Przeciętne wartości RMR dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnej liczby czynników (NOF) i rozpoznania typu modelu (jedno i niejednowymiarowe) z użyciem metod automatycznych (AOS)



Rysunek 42. Przeciętne wartości SRMR dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnej liczby czynników (NOF) i rozpoznania typu modelu (jedno i niejednowymiarowe) z użyciem metod automatycznych (AOS)

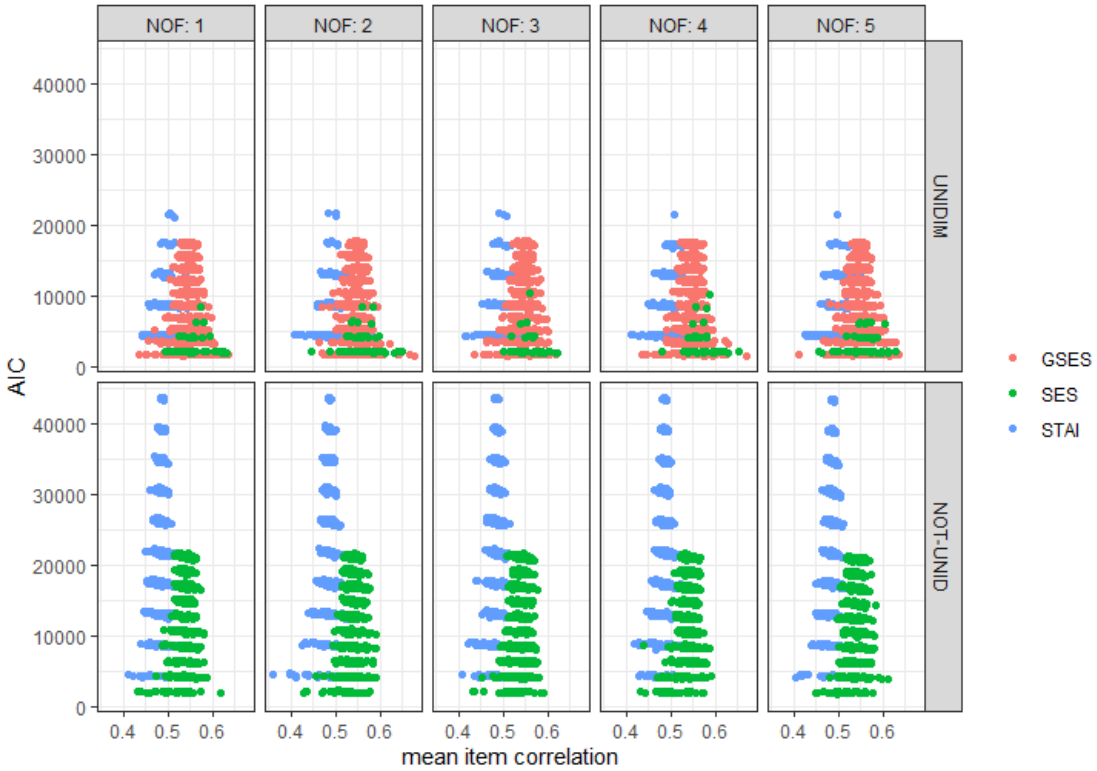


4.2.1.2. Miary przyrostowe

W dalszej kolejności zostaną przedstawione wyniki miar różnic między modelami. Graficzna prezentacja wyników została przygotowana analogicznie do poprzednich przykładów miar dopasowania – rysunki zbiorczo przedstawiają poziom danej wartości wskaźnika w podziale na rodzaj modeli (jedno- vs wielowymiarowe), liczbę czynników latentnych, przeciętny poziom korelacji i typ narzędzia badawczego.

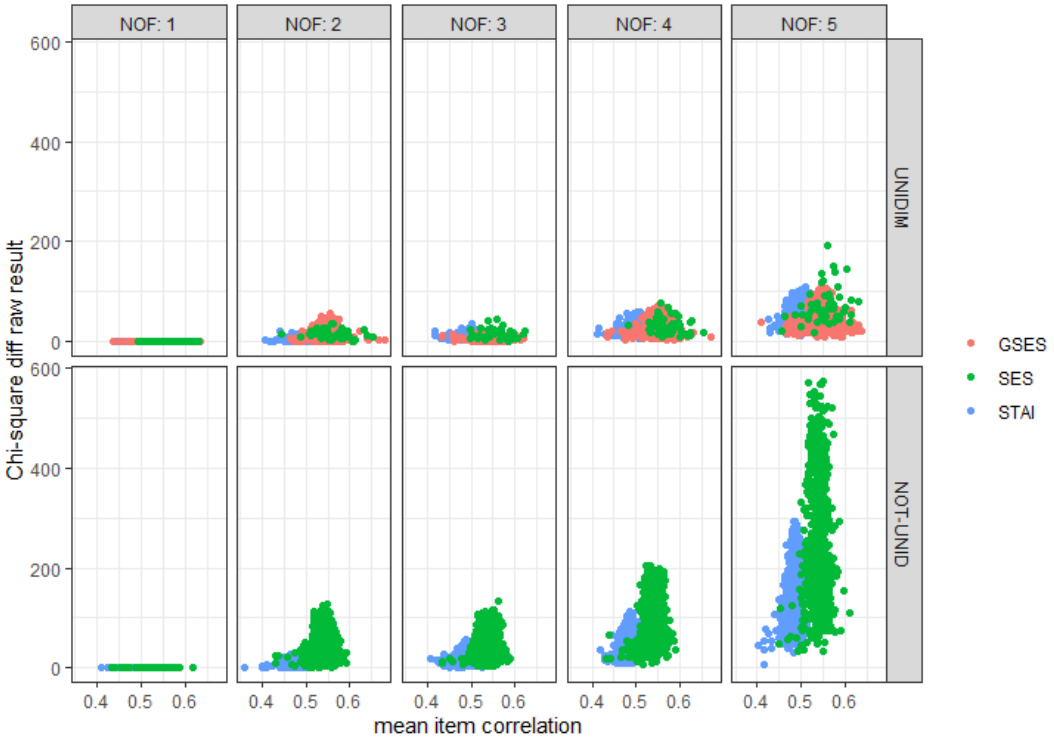
Miary AIC i BIC również były ze sobą silnie skorelowane, dlatego zostaną przestawione wyniki tylko dla AIC (Rysunek 43). Ponownie, zgodnie z wynikami uzyskanymi z symulacji, wartości współczynnika AIC uzyskiwane dla modeli nie różniły się pomiędzy strukturami teoretycznymi dla jednej, dwóch czy pięciu zmiennych latentnych, biorąc oczywiście pod uwagę rozpoznanie struktury (jedno- lub niejednowymiarowa); można było z taką samą szansą dopasować model jednoczynnikowy jak jego wielowymiarowe odpowiedniki. Interesująca jest kwestia rozpoznania modeli – modele rozpoznane jako niejednowymiarowe (niższy wiersz na Rysunek 43) miały po prostu większy poziom wartości AIC niż modele rozpoznane jako jednowymiarowe (wyższy wiersz na Rysunek 43). A więc można przyjąć, że po prostu dane, które są wielowymiarowe (lub istnieją wątpliwości co do ich jednowymiarowości) będą uzyskiwać przeciętnie większe wartości AIC niż dane jednowymiarowe. Nadal nie rozstrzyga to o dopasowaniu danych względem poprawnego modelu teoretycznego, jest jedynie problemem struktury danych: badacz nadal nie może ocenić na postawie AIC, czy w jego danych lepszy (prawidłowy) jest model jedno- czy wielowymiarowy, bo średnie wyniki uzyska zbliżone.

Rysunek 43. Przeciętne wartości AIC dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnej liczby czynników (NOF) i rozpoznania typu modelu (jedno i niejednowymiarowe) z użyciem metod automatycznych (AOS)



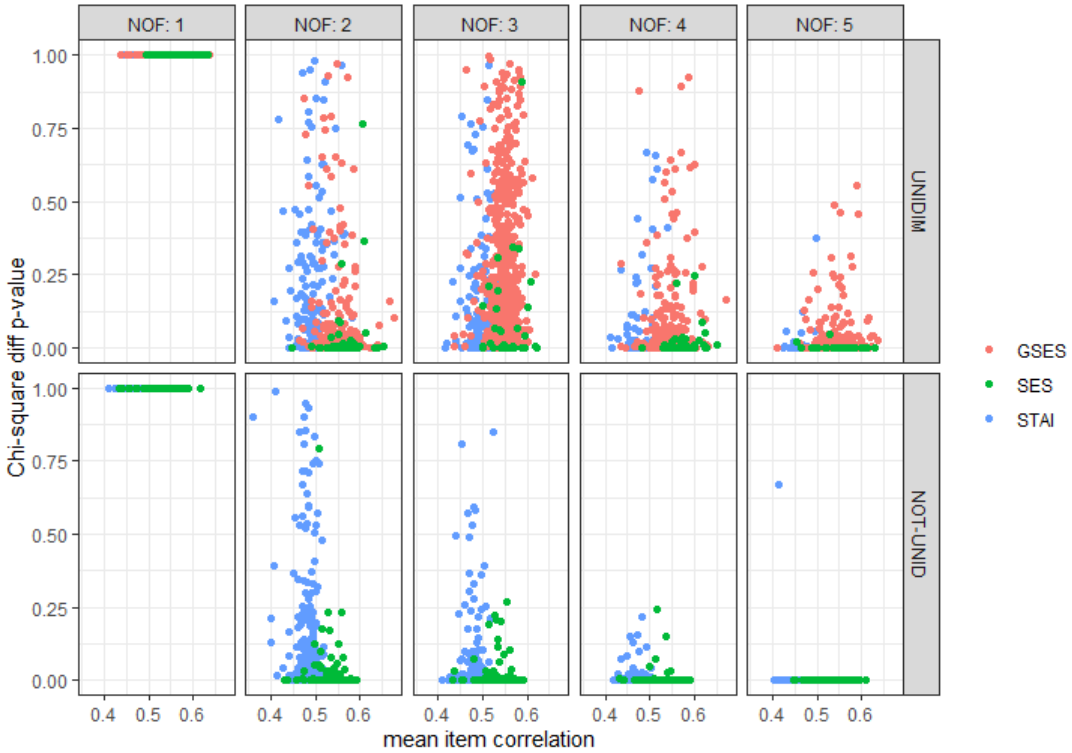
Jako następne przedstawione zostaną wyniki dla wartości surowej różnicy chi-kwadrat. Jeśli chodzi o miarę χ^2_{diff} (Rysunek 44), która może określać różnicę między modelem jednowymiarowym (poprawnym) a modelem wielowymiarowym (nieprawidłowym) to (jak w badaniu symulacyjnym) rzeczywiście rosła ona wraz z liczbą zmiennych latentnych w modelu, przy czym widoczny przyrost w macierzach danych o strukturze jednowymiarowej pojawia się dopiero w przypadku próby dopasowania modelu cztero- i pięciowymiarowego, a modele dwu- i trzywymiarowe mają zbliżone i często akceptowalne dopasowanie. Należy zwrócić uwagę na dwie kwestie. Po pierwsze, tak samo jak w badaniu symulacyjnym różnica w obrębie modeli o jednowymiarowej strukturze teoretycznej wynosi 0, co oznacza, że model jednowymiarowy nie różni się od sam od siebie - stąd na wykresach pozioma „kreska" na wartości 0 osi Y w tej grupie modeli. Po drugie, tak samo jak w badaniu symulacyjnym dla struktury dopasowywanej jako wielowymiarowa była obliczana wartość tych dwóch współczynników w oparciu o tworzony w locie kolejny, dodatkowy, jednowymiarowy model alternatywny oparty o te same dane. A więc różnica tak samo jak w przypadku badania symulacyjnego wyraża różnicę modelu wielowymiarowego (np. dwuczynnikowego) od modelu jednowymiarowego (dla tych samych danych).

Rysunek 44. Średnie wartości surowej różnicy chi-kwadrat (χ^2_{diff}) dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnej liczby czynników (NOF) i rozpoznania typu modelu (jedno i niejednowymiarowe) z użyciem metod automatycznych (AOS).



W przypadku modeli o strukturze danych niejednowymiarowej przyrosty są dużo większe i modele cztero- i pięciowymiarowe mogłyby zostać uznane za niedopasowane. Jednak biorąc pod uwagę istotność statystyczną $\chi^2_{diff\text{-}p}$ (Rysunek 45) dla modeli opracowanych na danych o strukturze jednowymiarowej okazywało się, że rozrzut tych wartości dla dwóch i trzech czynników jest tak duży, że różnice dla modeli są właściwie losowe i można z takim samym prawdopodobieństwem uznać model za poprawny jak i niepoprawny dla jednego, dwóch czy trzech czynników. W obrębie czwartego czynnika szanse na błąd maleją, ale nadal nie są niskie i dopiero dla czynnika piątego pojawia się duża szansa na uznanie struktury za nieprawidłową. Innymi słowy, rozbieżność między modelem teoretycznym a realną strukturą danych musiałaby być bardzo duża. W przypadku modeli o charakterze niejednowymiarowym szanse uzyskania istotnej różnicy wzrastają wraz z liczbą teoretycznych czynników w modelu, ale jeśli są to struktury o charakterze niejednowymiarowym, to prawdopodobnie kwestia tkwi w zasadzie jedynie w tym, że dane nie są jednowymiarowe w sposób widoczny i łatwiej dopasować model wielowymiarowy na danych wielowymiarowych. Z pewnością określenie z użyciem tych miar poprawności modelu jedno, dwu i trójwymiarowego nie jest w praktyce możliwe.

Rysunek 45. Poziom istotności dla różnicy chi-kwadrat ($\chi^2_{diff\text{-}p}$) dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnej liczby czynników (NOF) i rozpoznania typu modelu (jedno i niejednowymiarowe) z użyciem metod automatycznych (AOS)



Podsumowując wyniki dla miar klasycznych, przedstawione poziomy współczynników dopasowania dla struktur rozpoznanych jako jednowymiarowe właściwie nie różniły się od wyników badania symulacyjnego – poziom dopasowania modelu z użyciem klasycznych miar dopasowania był w ich obrębie właściwie taki sam, niezależnie od tego, ile zmiennych latentnych próbowano wyspecyfikować. Co ciekawe, w modelach rozpoznanych jako wielowymiarowe niewielką rolę grała średnia korelacja między pozycjami, czego nie zaobserwowano dla struktur (danych w obrębie punktów na wykresach) o charakterze niewątpliwie jednowymiarowym.

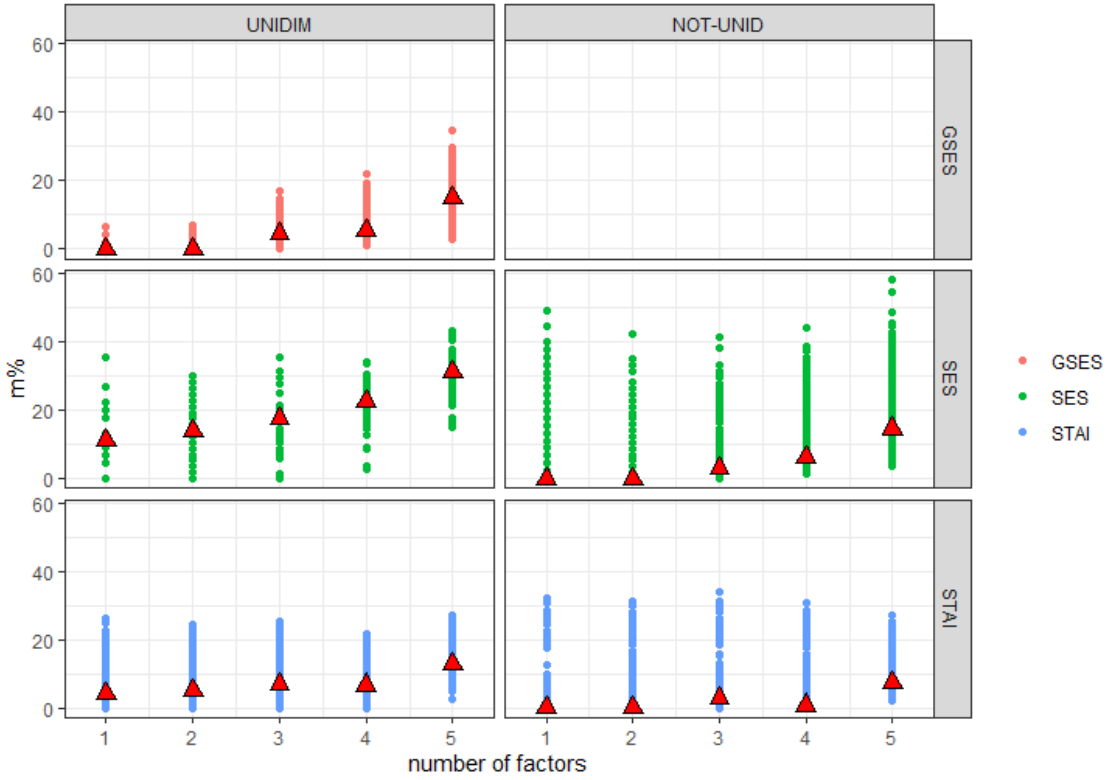
W grupie danych rozpoznawanych jako niejednowymiarowe wraz ze wzrostem liczby czynników spadała liczba modeli niedopasowanych (rzadziej zdarzały się punkty danych o słabych współczynnikach dopasowania), ale należy zauważyć, że po prostu mogło się w tej grupie modeli częściej zdarzać, że wylosowany i niejednowymiarowy zbiór danych miał tyle samo czynników, co ustalona losowo struktura teoretyczna. Ponadto warto zauważyć, nie tyle poprawiały się wszystkie wartości możliwych do uzyskania poziomów danego współczynnika dopasowania (w danej grupie modeli), co zmniejszała się liczba przekraczających próg akceptowalności, a więc zmieniała się wariancja zmiennej, natomiast wartości

maksymalne/minimalne poniżej progu nie ulegały zmianie (co widać na przykład na Rysunek 42).

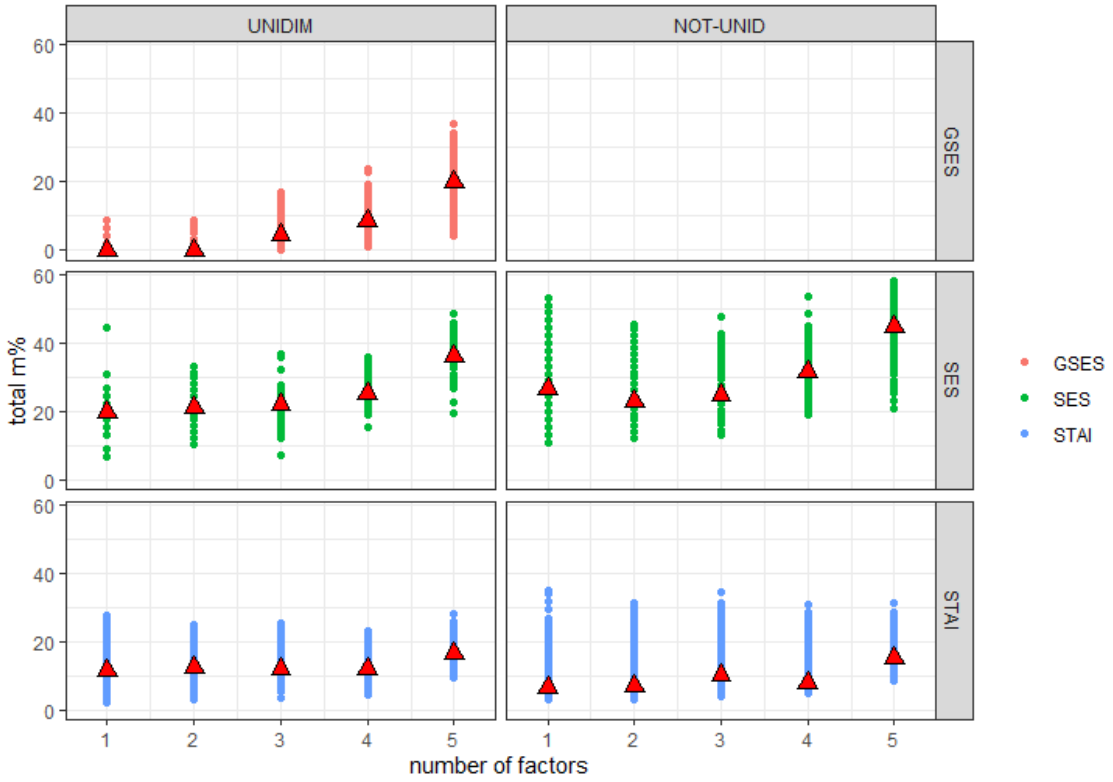
4.2.1.3. Indeksy modyfikacji

Jako ostatnie zostaną zaprezentowane wyniki dla miar opartych o indeksy dopasowania. Ich zestaw i metoda uzyskania (obliczenia) był identyczny jak w badaniu symulacyjnym. Silnie związane ze sobą parametry *m%* (Rysunek 46) i *total m%* (Rysunek 47) wzrastały wraz z liczbą czynników, ale między modelami jedno- a dwuwymiarowymi przyrosty były pomijalne, co wyklucza ich rozróżnienie. Parametr *epc:m%* nie pozwalał na rozróżnienie poszczególnych modeli (Rysunek 48), co wyklucza jego praktyczne zastosowanie.

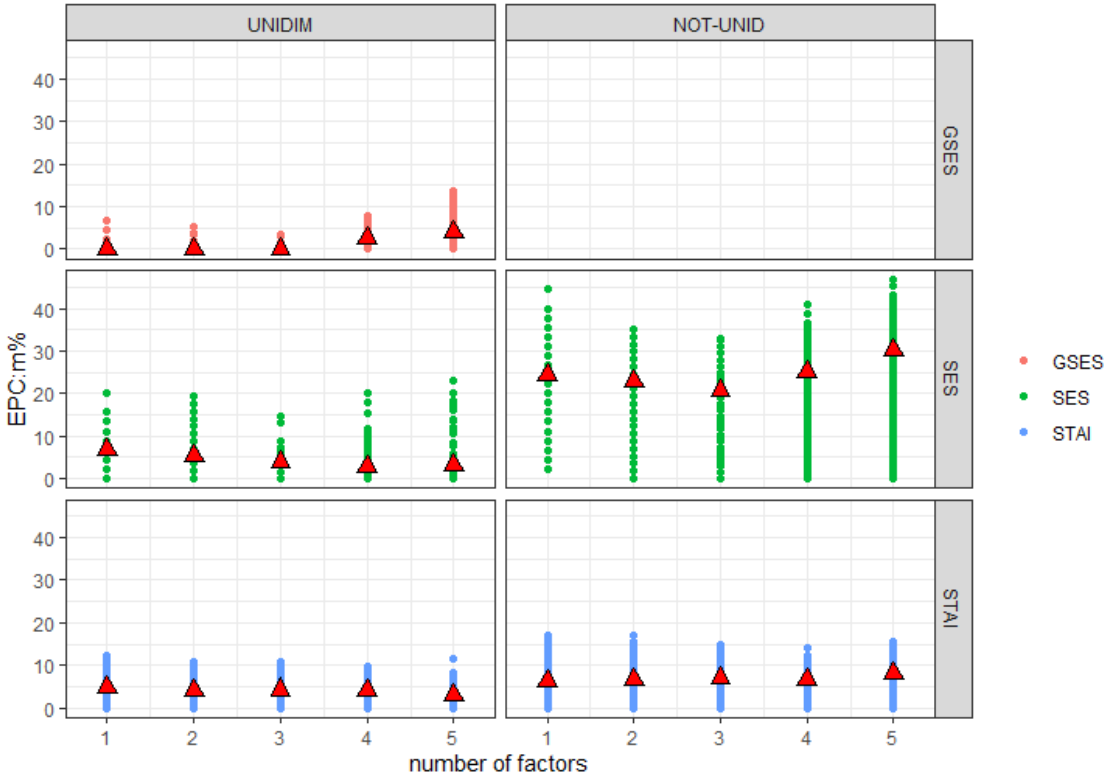
Rysunek 46. Przeciętne wartości m% dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnej liczby czynników (NOF) i rozpoznania typu modelu (jedno i niejednowymiarowe) z użyciem metod automatycznych (AOS)



Rysunek 47. Przeciętne wartości *total m%* dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnej liczby czynników (NOF) i rozpoznania typu modelu (jedno i niejednowymiarowe) z użyciem metod automatycznych (AOS)

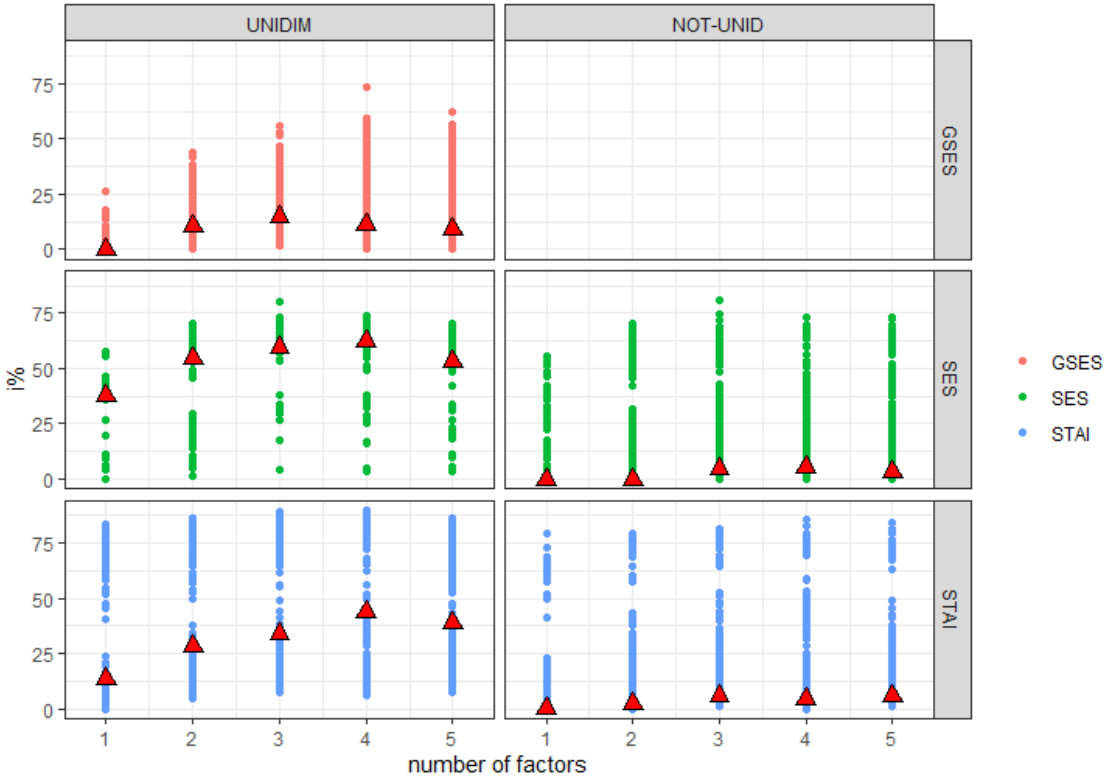


Rysunek 48. Przeciętne wartości *EPC:m%* dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnej liczby czynników (NOF) i rozpoznania typu modelu (jedno i niejednowymiarowe) z użyciem metod automatycznych (AOS)

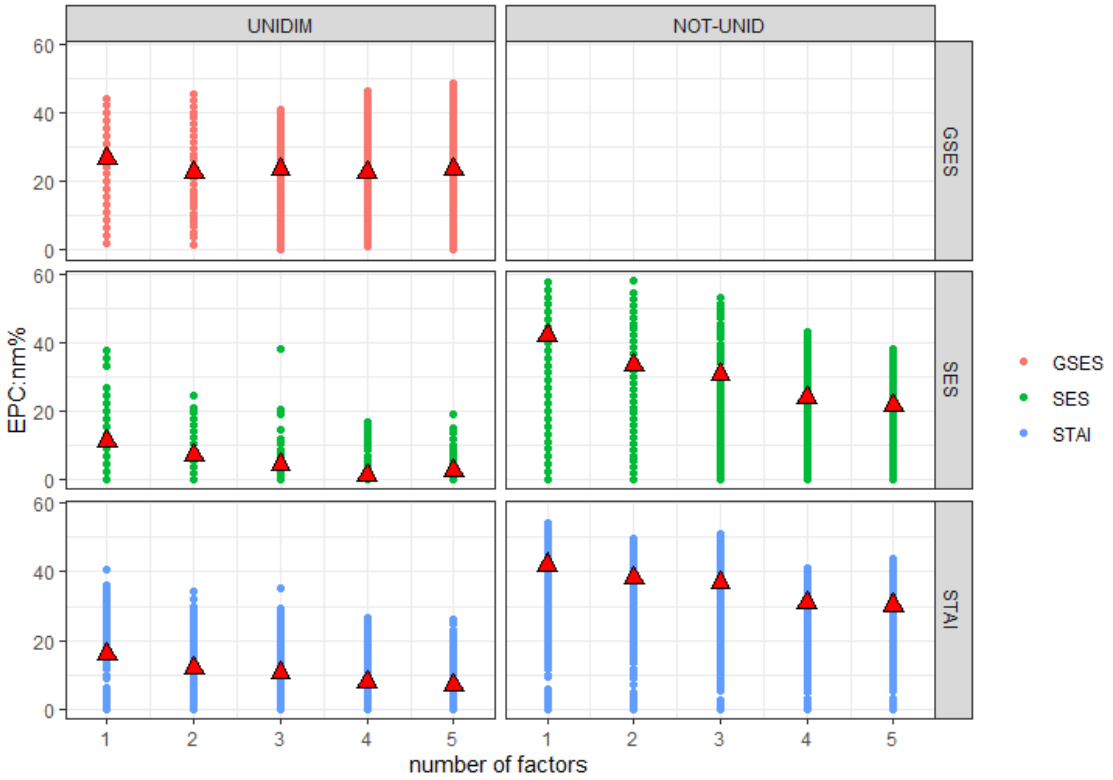


Liczba niekonkluzywnych ścieżek dla indeksów modyfikacji *i%* (Rysunek 49) wzrastała w danych rozpoznanych jako jednowymiarowe, ale przyrosty te były różne w obrębie trzech kwestionariuszy i, co ważne dla praktycznego zastosowania tego wskaźnika, przyrosty nie zawsze były jednoznaczne powyżej dwóch czynników. W obrębie badania symulacyjnego przyrosty były proporcjonalnie zależne od długości skali, ale w przypadku danych empirycznych takiej zależności nie zaobserwowano. Rozbieżność między badaniem symulacyjnym a badaniem empirycznym podważa wykorzystanie tego wskaźnika, choć wydaje się on obiecujący, ponieważ dla danych o strukturze niejednowymiarowej przyrostów tego wskaźnika nie było. Wartości współczynnika *EPC:nm%* (Rysunek 50) nie pozwalały w ogóle na żadną konkluzję, ponieważ spadki tej wartości były różne dla różnych narzędzi badawczych (różnych struktur empirycznych).

Rysunek 49. Przeciętne wartości *i%* dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnej liczby czynników (NOF) i rozpoznania typu modelu (jedno i niejednowymiarowe) z użyciem metod automatycznych (AOS)

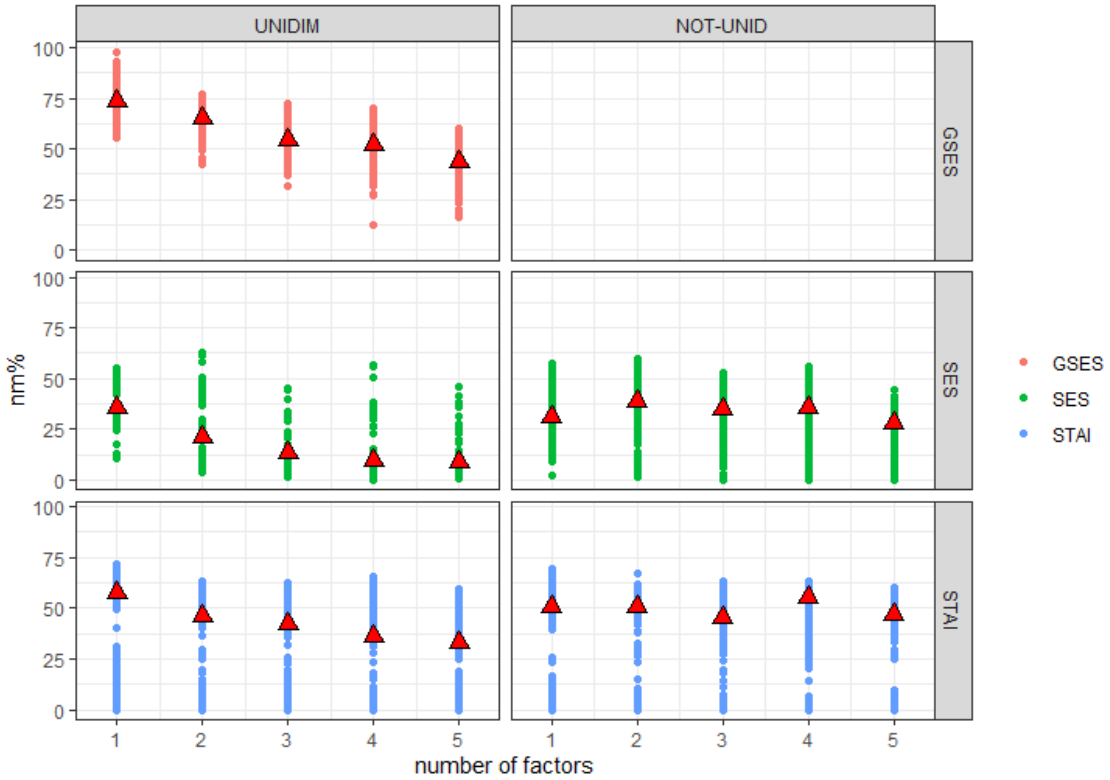


Rysunek 50. Przeciętne wartości *EPC:nm%* dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnej liczby czynników (NOF) i rozpoznania typu modelu (jedno i niejednowymiarowe) z użyciem metod automatycznych (AOS)



Dla miary *nm%* uzyskano zgodny z badaniem symulacyjnym poziom konkluzywności co do poprawnego rozpoznania struktury. W obrębie danych o charakterze jednowymiarowym, dla wszystkich trzech kwestionariuszy, widać (Rysunek 51) wyraźny spadek wartości odsetka nie-błędnie wyspecyfikowanych ścieżek w modelu, przy czym struktury nieprawidłowe uzyskują niższy poziom tego wskaźnika niż prawidłowe (jednowymiarowe). Dla danych o charakterze niejednowymiarowym przeciętne wartości *nm%* utrzymują się na zbliżonym poziomie dla wszystkich możliwych struktur teoretycznych. Oznacza to także, że miara może być z powodzeniem wykorzystywana w pośredniej detekcji właśnie struktur jednowymiarowych, ponieważ dla danych jednowymiarowych wraz z liczbą czynników spada jej wartość, a dla danych niejednowymiarowych średnia wartość jest niezależna od liczby czynników. A więc w przypadku, gdy struktura jest niewątpliwie jednowymiarowa, porównywanie modeli teoretycznych poprawnych i niepoprawnych może być wykonywane na podstawie tej wartości.

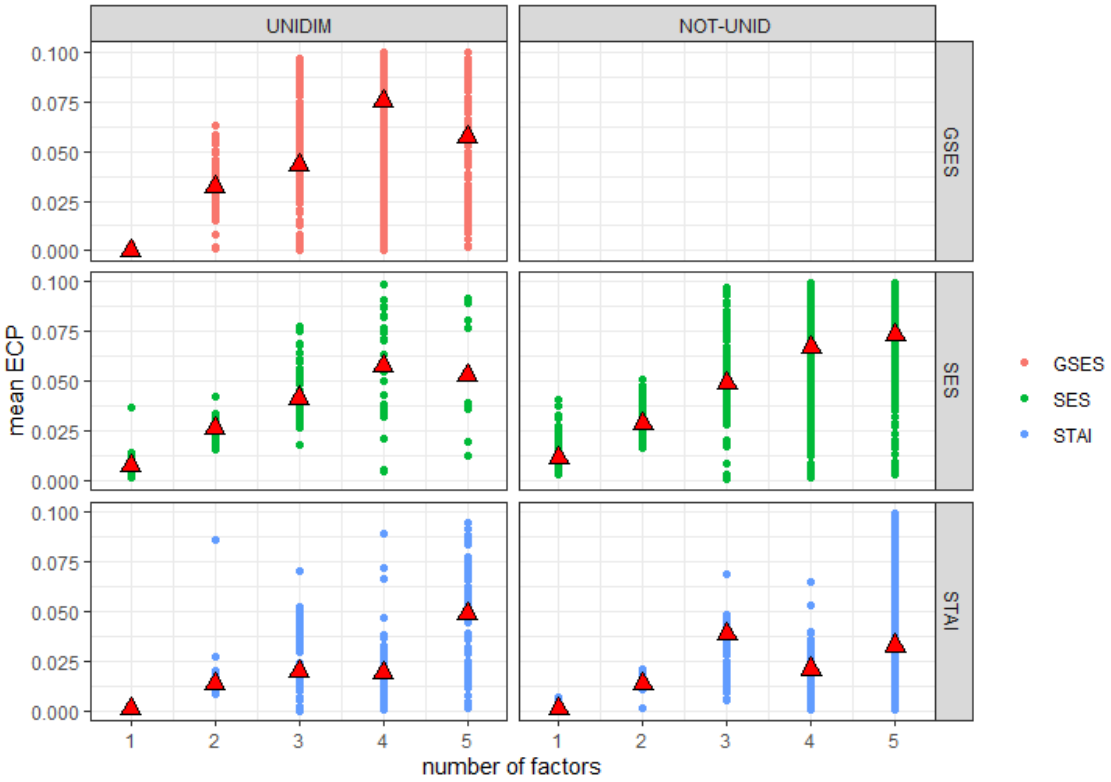
Rysunek 51. Wartości nm% dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnej liczby czynników (NOF) i rozpoznania typu modelu (jedno i niejednowymiarowe) z użyciem metod automatycznych (AOS)



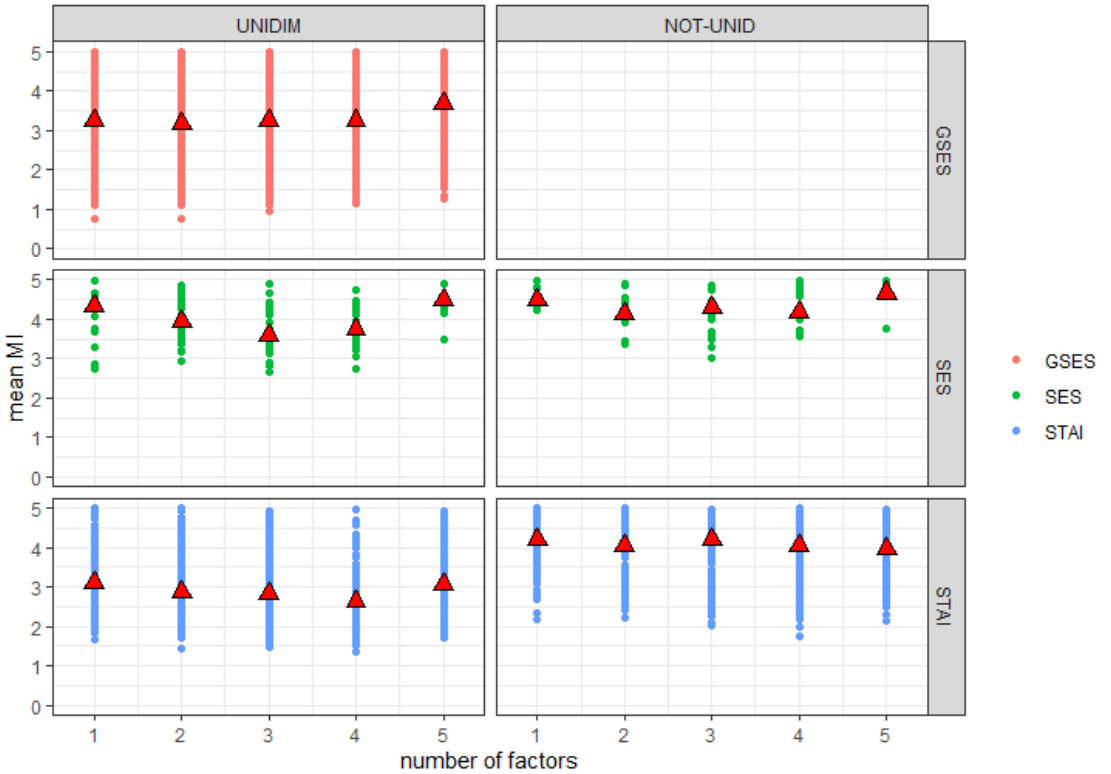
Przyrost średniej wartości *ECP* (Rysunek 52) zgodnie z wynikami symulacji mógł teoretycznie rozstrzygać, czy dane mają strukturę jednowymiarową, ale przyrost ten był także obserwowany w obrębie struktur rozpoznanych jako niejednowymiarowe. Trudno w takim razie ocenić, czy jest to kwestia samej miary *ECP*, która po prostu przyrasta wraz z liczbą czynników w modelu, czy raczej kwestia tego, że modele niejednowymiarowe uzyskane w badaniu empirycznym są po prostu zanieczyszczonymi modelami jednowymiarowymi czy modelami w istocie niejednowymiarowymi (zob. Rysunek 35). Finalnie, analogicznie do wyników symulacji, średnie wartości parametrów *MI* (Rysunek 53) i *NCP* (Rysunek 54) nie pozwalają na rozróżnienie danych jedno- i wielowymiarowych w żaden sposób – nie przyrastały ani nie spadały wyraźnie dla żadnego typu danych empirycznych.

Biorąc pod uwagę pytanie badawcze P2, zgodnie z badaniem symulacyjnym właściwie wszystkie klasyczne miary dopasowania okazały się nieprzydatne w określeniu, czy poprawny jest model jedno- czy wielowymiarowy, miary oparte o indeksy modyfikacji miały w tym zakresie większy potencjał przy wskazaniu miary *nm%* jako w najwyższym stopniu konkluzywnej.

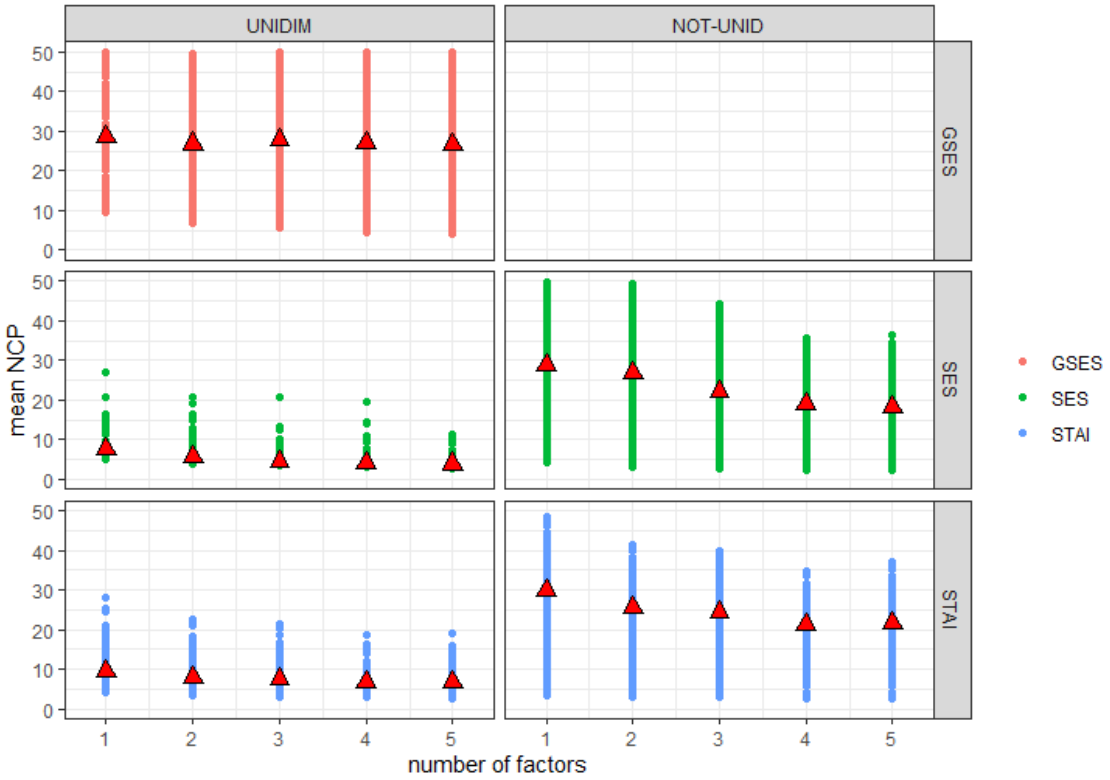
Rysunek 52. Wartości przeciętne parametru ECP dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnej liczby czynników (NOF) i rozpoznania typu modelu (jedno i niejednowymiarowe) z użyciem metod automatycznych (AOS)



Rysunek 53. Wartości przeciętne parametru MI dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnej liczby czynników (NOF) i rozpoznania typu modelu (jedno i niejednowymiarowe) z użyciem metod automatycznych (AOS)



Rysunek 54. Wartości przeciętne parametru NCP dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnej liczby czynników (NOF) i rozpoznania typu modelu (jedno i niejednowymiarowe) z użyciem metod automatycznych (AOS)



5. NIEEFEKTYWNA DROGA ROZRÓŻNIENIA PRZEZ PUNKTY ODCIĘCIA

Z uwagi na przedstawione w Tabela 4 potencjalne różnice między dwoma grupami modeli, postanowiono zbadać i znaleźć punkty odcięcia optymalne dla poszczególnych zmiennych, dla których były istotne statystycznie różnice, a tym samym odnieść się do pytania badawczego (P3) czy da się wskazać punkty odcięcia dla oceny, który model jest rzeczywiście jednowymiarowy. Analiza różnic przedstawiona w tabeli dotyczyła wszystkich analizowanych współczynników dopasowania modeli (zarówno globalnych miar, jak i lokalnych miar opartych o indeksy modyfikacji), a istotne różnice pojawiły się dla miar globalnych SRMR i χ^2 oraz wszystkich miar lokalnych. Warto przy tym ponownie zwrócić uwagę na kwestię tego, iż uzyskane istotne różnice dla miar globalnych SRMR i χ^2 były prawdopodobnie wynikiem analizy na bardzo dużych próbach, ponieważ miary siły efektu były bardzo małe, a różnice między wartościami dla modeli poprawnych i niepoprawnych właściwie nierozróżnialne.

Jeśli różnica między poprawnymi (jednowymiarowymi) a niepoprawnymi (wielowymiarowymi) modelami byłaby wystarczająco duża, by pomiędzy nimi umieścić próg wartości diagnostycznej, analiza ROC (*receiver operating characteristic*, Unal, 2017) wykazałby to, wskazując optymalny punkt odcięcia pomiędzy klasami modeli. Niestety, niezależnie od użytej metody obliczeń (analiza ROC może zostać przeprowadzona z użyciem kilku sposobów obliczeń dla których wynik jest zwykle zgodny, Unal, 2017) wyniki nie pozwalały na wyciągnięcie sensownych wniosków. Z uwagi na tzw. wiązania (powtarzające się takie same wartości zmiennej zależnej w obu porównywanych grupach modeli) i niewielkie różnice między modelami optymalne punkty odcięcia nie były aplikowalne w sposób jednoznaczny i praktyczny. Jak widać w Tabela 9, punkty odcięcia wypadały dla wartości, które nie znajdowały się w przedziałach wyników średnich, lub różnice między wartościami były bardzo małe. Oznacza to, że biorąc pod uwagę możliwy do uzyskania rozrzut wyników dla wszystkich wymienionych współczynników podział modeli na poprawne i niepoprawne dałby tu jedynie przypadkową wartość punktu odcięcia. Zatem i ta droga nie daje satysfakcjonujących wyników. Dodatkowo, biorąc pod uwagę małe siły efektu różnic międzygrupowych, można z dużą dozą pewności uznać, że przynajmniej część tych różnic wynika jedynie z wielkości próby.

Wartości o charakterze nie pozwalającym na uzyskanie interpretowalnych rezultatów pojawiły się dla miar *epc:nm%*, *epc:m%*, oraz przeciętnej wartości parametru *NCP*. Zbyt małe wartości lub wymagające rozróżnienia na trzecim lub czwartym miejscu po przecinku

dotyczyły *i%*, *total m%* i przeciętnej wartości parametru *ECP*. Wyniki analiz ROC w zakresie tych współczynników należy traktować jako w najlepszym wypadku niepewne.

Tabela 9
Próba wyznaczenia punktów odcięcia wartości współczynników rozróżniających modele jedno- i wielowymiarowe metodą ROC

	Mdn_{multi}	Mdn_{uni}	Powierzchnia	p	MAX	punkt odcięcia
χ^2	443,89	448,30	0,51	0,013	1,05	34,63
SRMR	0,0236	0,0242	0,51	0,016	1,02	0,0206
nm%	76,20	93,33	0,78	< 0,001	1,58	91,02
m%	1,67	0,00	0,23	< 0,001	1,00	0,069
epc:nm%	5,37	5,31	0,47	< 0,001	1,03	4,44
epc:m%	0,00	0,00	0,42	< 0,001	1,00	0,069
i%	10,38	0,00	0,24	< 0,001	1,05	0,001
total m%	2,15	0,00	0,21	< 0,001	1,00	0,001
MI_M	1,27	1,07	0,30	< 0,001	1,33	1,24
ECP_M	0,00	0,00	0,35	< 0,001	1,33	0,001
NCP_M	39,97	51,37	0,53	< 0,001	1,05	130,88
α	0,847	0,949	0,73	< 0,001	1,33	0,901

Interesujące wyniki uzyskano dla miar *nm%*, *m%* i przeciętnej wartości *MI*. Analiza w pierwszej kolejności wskazywałaby, że można z użyciem tych miar wyznaczyć w dość racjonalne progi odcięcia, ale – jak to przedstawiono wcześniej – miary te (a właściwie różnice w tych miarach) nie powinny być wykorzystywane dla dowolnego typu danych (na przykład skal użytych w badaniu). Na Rysunek 27 przedstawiono, jak zachowuje się miara n*m*% w zależności od typu skali Likerta użytej w badaniu – miara ta spada wraz z kolejnymi dodanymi nieprawidłowo wymiarami latentnymi, ale w zależności od użytego typu skali jej maksymalne i minimalne wartości są inne. Wyznaczenie zatem progu odcięcia w przypadku tej miary jest o tyle niepraktyczne, że należałoby wyznaczyć inne progi odcięcia dla różnych typów tworzonych narzędzi badawczych, a liczba różnych typów narzędzi badawczych (długości skal, liczba wymiarów, ich kombinacji) z perspektywy przytoczonych wyników w żadnym wypadku nie umożliwi ujednoznacznienia wytycznych w użyciu progów odcięcia. W przypadku miar *m%* i przeciętnego parametru *MI* można mówić o generalnych różnicach między modelami prawidłowymi i nieprawidłowymi, ale obie te miary (jak wskazano przy

omawianiu Rysunek 23 i Rysunek 32) nie pozwalają na prawidłowe rozróżnienie modeli jedno- i dwuwymiarowych, zatem i dla nich wyznaczenie wartości progowej nie daje wyników potrzebnych do określenia prawidłowo struktur latentnych o jednowymiarowym charakterze.

Odpowiadając na pytanie badawcze P3, należy stwierdzić, że nie jest możliwe wyznaczenie użytecznych empirycznie punktów podziału (punktów odcięcia, *rule of thumb*), które pozwalałyby na trafną ocenę jedno- lub wielowymiarowości modelu, ponieważ analiza krzywej ROC nie doprowadziła do identyfikacji wartości granicznych o wystarczającej wartości diagnostycznej.

6. DYSKUSJA

6.1 Podsumowanie i dyskusja uzyskanych wyników

Jednym z kluczowych wyzwań w psychometrycznej ocenie narzędzi pomiarowych jest trafne określenie ich struktury wymiarowej – a więc rozstrzygnięcie, czy dany konstrukt jest jedno- czy wielowymiarowy. Konfirmacyjna analiza czynnikowa (CFA) stanowi standardowe narzędzie służące do weryfikacji przyjętych modeli teoretycznych, opierając się na wskaźnikach dopasowania, które mają pomóc w ocenie, na ile dany model odpowiada rzeczywistym danym. Jednakże, pomimo powszechnego stosowania tych miar, pojawiają się uzasadnione wątpliwości, czy klasyczne (globalne) miary dopasowania rzeczywiście umożliwiają jednoznaczne rozróżnianie poprawnych i niepoprawnych modeli, zwłaszcza w przypadkach danych o jednoznacznie jednowymiarowej strukturze. Celem niniejszej pracy było zatem krytyczne zweryfikowanie przydatności klasycznych współczynników dopasowania w tego typu sytuacjach oraz poszukiwanie alternatywnych rozwiązań. W tym kontekście postawiono trzy pytania badawcze dotyczące (P1) zdolności miar dopasowania do rozróżniania modeli jedno- i wielowymiarowych, (P2) trafności poszczególnych wskaźników oraz (P3) możliwości ustalenia praktycznych punktów odcięcia. Aby udzielić na nie odpowiedzi, przeprowadzono dwa uzupełniające się badania: rozbudowane badanie symulacyjne z wykorzystaniem jednowymiarowych danych oraz badanie empiryczne na dużej próbie rzeczywistych respondentów. Wyniki tych analiz stanowią podstawę do dalszych rozważań przedstawionych poniżej.

W obrębie symulacji (badania symulacyjnego), jako pierwsze analizowano różnice w średnim poziomie wszystkich wskaźników dopasowania między modelami prawidłowymi i nieprawidłowymi. Innymi słowy, dokonano porównania uśrednionych miar dopasowania struktur jednowymiarowych oraz prób dopasowania tych samych struktur jako wielowymiarowe w odpowiedzi na pytanie badawcze P1. Jeśli chodzi o klasyczne (globalne) miary dopasowania modelu, to uzyskano istotne statystycznie różnice jedynie dla surowej wartości ogólnej modelu χ^2 i SRMR, ale miały one bardzo małe siły efektu, a różnice pomiędzy typami modeli były właściwie niezauważalne W przypadku obu analizowanych współczynników, ze względu na znaczny rozrzut możliwych wartości, ich wykorzystanie do jednoznacznej identyfikacji poprawnego modelu okazuje się w praktyce niewiarygodne. Niezależnie od tego, czy model odzwierciedlał rzeczywistą strukturę danych, czy też nie,

zbliżone prawdopodobieństwo uzyskania wartości sugerujących dobre dopasowanie sprawiało, że interpretacja tych wskaźników była niejednoznaczna i mało diagnostyczna.

Istotne i wyraźne różnice (większe siły efektu) zostały zaobserwowane dla wszystkich współczynników opartych o indeksy modyfikacji. Wyniki wskazują przede wszystkim, że nieprawidłowe modele miały wyraźnie różne od modeli poprawnych wartości indeksów podstawowych (*MI*, *ECP* i *NCP*) i szczegółowych (*m, nm*, *EPC:m*, *EPC:nm* lub *i*). Oznaczało to, już na tym etapie, że – względnie rzadko stosowane indeksy modyfikacji – mogą mieć spory potencjał w wykrywaniu poprawnej liczby wymiarów danego narzędzia, w przeciwieństwie do klasycznych miar dopasowania, stosowanych obecnie dość powszechnie w psychometrii.

W każdym przypadku, zarówno dla modeli poprawnych (jednowymiarowych) jak i niepoprawnych (wielowymiarowych), wartości klasycznych indeksów dopasowania przekraczały przeciętnie progi pozwalające na uznanie struktury za dopasowaną. Oznacza to, że samo przekroczenie sugerowanego obecnie przez literaturę danego progu akceptacji, nie rozstrzygało o tym, który model (jedno- czy wielowymiarowy) badacz powinien wybrać i z praktycznego punktu widzenia znane reguły odcięcia nie były przydatne do oceny któregoś modelu jako lepszy. Z uwagi na to, postanowiono poszukać nowych punktów odcięcia optymalnych dla poruszonego tu problemu badawczego P3. Niestety, wyniki analizy ROC nie pozwalały na wyciągnięcie jednoznacznych wniosków. Niewielkie różnice między modelami uniemożliwiały znalezienie praktycznych reguł odcięcia w obrębie grupy modeli jednowymiarowych.

W dalszej kolejności przeprowadzono przegląd parametrów modeli pod względem problemu badawczego P2, bardziej szczegółowo sprawdzając, jak wspólny wpływ różnych warunków badawczych (na przykład średniej korelacji między pozycjami i liczby czynników) może zmieniać poszczególne parametry dopasowania. Okazało się, że niezależnie od tego, jakie parametry miały testowane dane jednowymiarowe – na przykład to, jaką miały liczbę itemów w kwestionariuszu, lub jakie były średnie związki między tymi itemami oraz ich różne kombinacje – nadal rozróżnienie modeli jedno- i wielowymiarowych opartych o dane jednowymiarowe po prostu nie było możliwe. Tendencje centralne w tych wynikach są praktycznie nierozróżnialne. Także rodzaj rozkładu (normalny vs jednorodny), długość skal Likerta czy wielkość próby nie miały wpływu na wyniki uzyskane dla wskaźników dopasowania. Dotyczyło to wszystkich klasycznych miar dopasowania, w tym także miar, których zadaniem jest faktyczne porównywanie modeli ze sobą.

Choć popularne miary dopasowania, okazały się nieszczególnie przydatne w określeniu poprawnej specyfikacji wymiarowości modelu, uzyskano spore i wyraźne różnice

dla miar opartych o indeksy modyfikacji. Miary te wyraźnie różniły modele jedno- i wielowymiarowe, ponadto zmieniały się wraz z liczbą nieprawidłowych wymiarów kwestionariusza, niemniej jednak większość z nich (*MI*, *NCP*, *ECP*, *m, epc:m, epc:nm, i*) nie pozwalało na rozróżnienie modeli jedno- i dwuwymiarowych, wyraźnie wskazując, że dopiero trójwymiarowy model jest błędny. Oznacza to, że badacz nie mógłby zdecydować w pewnej grupie przypadków, czy konstrukt jest w rzeczywistości jedno- czy dwuwymiarowy, mógłby jedynie odrzucać jako nieprawidłowe modele trzy- i więcej wymiarowe. Natomiast jedna z miar bardzo wyraźnie pozwalała na wykrycie nieprawidłowości modelu wielowymiarowego – wraz ze wzrostem liczby czynników wyraźnie spadają wartości *nm%,* przy czym spadek ten jest widoczny już od pierwszego niepoprawnie dodanego wymiaru konstruktu.

Podsumowując, właściwie wszystkie klasyczne miary dopasowania okazały się nieprzydatne w określeniu, czy poprawny jest model jedno- czy wielowymiarowy. Przydatne okazały się być wyniki oparte o indeksy modyfikacji, a decyzje oparte o miarę *nm%* pozwalają na najbardziej precyzyjne odrzucanie niepoprawnych modeli.

W obrębie badania empirycznego udało się potwierdzić właściwie wszystkie wyniki uzyskane w symulacjach. Klasyczne, globalne miary dopasowania nie pozwalały na rozróżnienie modeli CFA prawidłowych (jednowymiarowych) od nieprawidłowych (wielowymiarowych), a co za tym idzie utworzenia reguł odcięcia. Wszystkie wartości były zbliżone i szansa dopasowania modelu jako „poprawnie" jedno- lub wielowymiarowego była właściwie taka sama. Podobnie zatem jak w badaniu symulacyjnym, jednoznaczne ustalenie, czy struktura jest jedno- czy wielowymiarowa z użyciem tych miar jest niemożliwe, bo można je tak samo dobrze (lub akceptowalnie) dopasować. Analizując wyniki bardziej szczegółowo, także tutaj rozpatrzono wszystkie warunki różniące wyniki: liczbę dopasowywanych czynników czy średnią uzyskaną korelację w obrębie kwestionariusza. Również tutaj okazało się, że tak jak w symulacjach, miary właściwie się nie różnią od siebie we wszystkich typach modeli – ani rozpoznanie typu modelu, ani liczba czynników nie miały tu wpływu na całkowicie losowe szanse dopasowania modeli. Wnioski dla pytań badawczych P2 i P3, które były kluczowymi problemami badania empirycznego, były całkowicie zgodne z wnioskami płynącymi z badania symulacyjnego.

Tak jak w badaniu symulacyjnym, udało się natomiast uzyskać wyraźne różnice dla miar opartych o indeksy dopasowania. Ponownie, wszystkie miary dopasowania z tej grupy silnie zmieniały się wraz z liczbą kolejnych, zbędnych w modelu teoretycznym wymiarów kwestionariusza, ale sporych kłopotów nastręczało rozróżnianie z ich użyciem modeli jedno- i dwuwymiarowych. Jedynie dla miary *nm%* uzyskano zgodny z badaniem symulacyjnym

poziom konkluzywności co do poprawnego rozpoznania struktury. W obrębie danych o charakterze jednowymiarowym, dla wszystkich trzech kwestionariuszy empirycznych (SES, STAI i GSES), uzyskano wyraźny spadek wartości tej miary dla ogółu wszystkich 15000 punktów danych. Nie tylko oznacza to, że symulacja dość realistycznie odzwierciedliła opracowywane tu zagadnienia, ale też, że przewidywania na nich oparte były poprawne.

Finalnie – właściwie wszystkie uzyskane wyniki w badaniu empirycznym były zbieżne z wynikami z badania symulacyjnego. Jedyną wyraźną różnicę odnotowano w przypadku współczynnika *EPC*. W ramach analizy symulacyjnej za wartość sugerującą potencjalną nieprawidłowość specyfikacji modelu uznano przeciętny poziom *EPC* przekraczający 0,010. Oznaczało to, że wyższe wartości mogły wskazywać na konieczność wprowadzenia dodatkowych powiązań między zmiennymi (a w efekcie uznania modelu za nieprawidłowy). W badaniu empirycznym zaobserwowano, że wzrost średniej wartości *EPC* – zgodnie z ustaleniami symulacyjnymi – mógłby teoretycznie świadczyć o jednowymiarowości struktury danych. Jednakże taki wzrost pojawiał się również w przypadkach, które zostały zaklasyfikowane jako struktury wielowymiarowe. Trudno zatem jednoznacznie ocenić, czy jest to ograniczenie samej miary *EPC*, która może systematycznie wzrastać wraz z liczbą czynników w modelu, czy raczej efekt tego, że modele wielowymiarowe zidentyfikowane w analizie empirycznej odzwierciedlają faktycznie złożone, lecz nadal częściowo jednowymiarowe konstrukty. Innymi słowy w obrębie badania empirycznego różnice dla *ECP* nie były tak jednoznaczne jak dla wyników badania symulacyjnego. Pojawia się więc wątpliwość, czy przewidywania symulacyjne dla tego współczynnika należy odrzucić (lub traktować z pewną dozą wątpliwości), czy też potrzebne są dalsze badania mogące rozstrzygnąć o powodach tych nieznacznych rozbieżności między wskaźnikami. Niniejsza praca tego nie rozstrzyga. Uzyskane różnice mogą wynikać z poziomu zanieczyszczenia danych źródłowych (niepewnych struktur czynnikowych), kwestii indywidualnych, które nie były badane (wpływ czynników socjodemograficznych, osobowościowych itd.), albo nieznanych błędów zbierania danych.

Biorąc powyższe pod uwagę, można zaproponować następujące odpowiedzi na sformułowane pytanie

- Na pytanie badawcze (P1) *czy rzeczywiście współczynniki dopasowania różnią się istotnie pomiędzy modelami jedno- i wielowymiarowymi utworzonymi na tych samych danych o jednowymiarowej strukturze* należy udzielić przeczącej odpowiedzi. Nie udało się znaleźć jednoznacznych różnic między takimi modelami i z perspektywy

przedstawionych tu wyników modele jedno- i wielowymiarowe oparte o dane jednowymiarowe nie różnią się średnimi wartościami współczynników dopasowania.

- Jeśli chodzi o pytanie badawcze (P2), *które miary dopasowania i oceny modelu CFA pozwalają na najbardziej jednoznaczne wnioski dotyczące akceptacji bądź odrzucenia modelu w przypadku, gdy konstrukt jest jednowymiarowy, ale badacz nie jest pewny ile ma on wymiarów* - należy odpowiedzieć, że jedynymi współczynnikami pozwalającymi na rozróżnienie takich modeli w analizie CFA są miary oparte o indeksy modyfikacji, z czego jedna z nich *nm%* jest miarą co do której wnioskowanie jest dość jednoznaczne.
- W obrębie pytania badawczego (P3), *czy da się wskazać punkty odcięcia dla oceny, który model jest rzeczywiście jednowymiarowy* należy odpowiedzieć, że wyznaczenie wiarygodnych i możliwych do wykorzystania w praktycznej pracy progów odcięcia okazało się niemożliwe.

Zbliżone wyniki do tu zaprezentowanych, używając podobnej metodologii uzyskał Finch (2020). Jego badanie jednak dotyczyło tylko analizy mocy RMSEA, CFI i TLI, nie uwzględniało skal Likerta (tylko liczby rzeczywiste), a ponadto badanie miało mniejszą liczbę przedziałów interkorelacji pozycji (0,20, 0,45, 0,70) i mniejszy zakres liczebności podgrup (100-500). Niemniej, wyniki w badanym przez niego zakresie były niemal identyczne z przedstawionymi w niniejszej pracy: RMSEA, TLI i CFI pomiędzy różnymi testowanymi modelami prawidłowymi i nieprawidłowymi prawie się nie różniły. Co warto zauważyć w świetle przedstawionych tu wyników, mimo że różnice były na trzecim lub czwartym miejscu po przecinku, autor wspomnianego badania uznał, że te niewielkie wartości różnic mogą posłużyć do zbudowania punktów odcięcia. Badanie Finch (2020) nie uwzględniało w żaden sposób rozkładu wyników i oparte tylko o symulacyjną metodologię mogło pominąć problem niejednoznaczności wyników uzyskiwanych w realnych badaniach. Przedstawione w niniejszej pracy przykładowe publikacje dotyczące walidacji narzędzi badawczych wskazują na to, że różnice między poszczególnymi typami modeli (poprawnie i niepoprawnie wyspecyfikowanymi) z praktycznego punktu widzenia często są zbyt małe, by w świetle rzeczywistych wyników badań i praktyki tworzenia narzędzi psychometrycznych uznać je za konkluzywne. Co za tym idzie podejście Fincha i wnioski płynące z jego artykułu można uznać za artefaktyczne, nie dostrzegające realnego prawdopodobieństwa uzyskania wyników zbyt zbliżonych lub niejednoznacznych, by wyznaczone w taki sposób punkty odcięcia mogły jednoznacznie orzekać o jedno- lub wielowymiarowości analizowanego narzędzia badawczego.

Pośrednio zgodne z uzyskanymi tu rezultatami są wyniki Garrido, Abad i Ponsoda (2016). W badaniu tym analizowano skuteczność różnych wskaźników dopasowania (użyto indeksów RMSEA, TLI, CFI i SRMR) w określaniu liczby czynników w analizie czynnikowej z wykorzystaniem zmiennych kategorialnych, a ich skuteczność porównano z (wcześniej opisaną) metodą analizy równoległej. Badacze wykazali, że różnice we współczynnikach dopasowania między modelami poprawnymi a niepoprawnymi podczas ustalania liczby wymiarów jakiegoś narzędzia są zwykle bardzo małe nawet w idealnych sytuacjach. Wyszczególnione przez badaczy były przede wszystkim trudności w ustaleniu optymalnych wartości granicznych dla wskaźników dopasowania oraz wyższa skuteczność analizy równoległej niż wskaźników dopasowania, wraz z sugestią, że badacze powinni uzupełniać swoje teoretyczne rozważania dotyczące liczby czynników o wyniki uzyskane za pomocą tej metody analizy równoległej. Choć jest to wniosek pośrednio zgodny z wynikami uzyskanymi w niniejszej pracy, należy nadmienić, że autorzy nie analizowali realistycznych danych, jedynie posługiwali się danymi symulacyjnymi o niezanieczyszczonym charakterze, a samo badanie symulacyjne dotyczyło zmiennych jakościowych, co znów niekoniecznie aplikuje się do rozważanych tu zastosowań w psychologii i psychometrii.

Brauer i Ranger (2021) po przeprowadzeniu swoich badań symulacyjnych konkludowali, że miara χ^2, a co za tym idzie inne miary dopasowania oparte o tę miarę (na przykład RMSEA), nie powinny mogą być wykorzystywane do poprawnego określania, czy teoretyczna liczba wymiarów jest poprawną specyfikacją modelu, ponieważ różnice w wartościach tych indeksów dopasowania są albo zbyt małe, albo ich zakresy wzajemnie na siebie nachodzą. Nadmienić jednak należy, że ich badania symulacyjne dotyczyły jedynie analiz o charakterze eksploracyjnym (obiektem zainteresowania wspomnianego artykułu były analizy EFA/PCA), a dla miar dopasowania w analizie CFA nie przeprowadzili właściwie żadnej symulacji i nie wskazali dowodu potwierdzającego empirycznie czy ich konkluzja jest prawidłowa.

Interesujące i względnie nowe (na luty 2025 roku zaledwie kilka artykułów tych samych dwóch autorów) podejście do progów odcięcia dla miar dopasowania, w celu wskazywania modeli jako poprawnie wyspecyfikowanych i niepoprawnie wyspecyfikowanych, pojawia się na przykład u McNeish i Wolf (2023). Badacze zwrócili uwagę w obrębie badań symulacyjnych, że pojedyncze progi odcięcia dla danej miary mogą błędnie oceniać model poprawny jako niepoprawny, dlatego zaproponowali oni alternatywne podejście poprzez tak zwane *Dynamic Fit Index Cutoffs* (podejście *DFIC*). Ich zdaniem, różne grupy modeli mogą mieć różne poziomy akceptowalnego dopasowania: na przykład dany próg RMSEA

(analizowano także SRMR i CFI) powinien być uznawany za pozwalający na akceptację modelu lub nie pod warunkiem uwzględnienia m. in. różnych aspektów samego modelu SEM (wielkość próby, liczba itemów, liczba czynników) oraz uzyskanych statystyk (rozkładów, albo jednoczesnej oceny kilku współczynników dopasowania). Obecnie jednak nie ma jasności co do tego, ile różnych aspektów modelu należy uwzględnić w podziale na podgrupy, aby był on wyczerpujący, a badaczom nie udało się uzyskać żadnego zestawu takich progów, który pozwalałby na całkowicie prawidłową ocenę każdego z symulowanych modeli, a procent poprawnie rozpoznanych modeli przy realistycznie dobranych wartościach symulacji (czyli np. przeciętne związki między pozycjami kwestionariusza) nie przekraczał sześćdziesięciu procent. Biorąc pod uwagę wszystkie przedstawione w niniejszej pracy wyniki, czyli podział poszczególnych wyników według wszystkich kontrolowanych zmiennych w symulacjach, należy postawić pytanie, czy podejście *DFIC* w ogóle pozwala na prawidłową ocenę modeli. Na koniec należy zauważyć, że a) samo badanie McNeish i Wolf (2023) nie dotyczyło kwestii rozstrzygania o jedno- lub wielowymiarowości a jedynie o poziomie dopasowania do danych modelu samego w sobie; b) nie zostało ono przeprowadzone na żadnych realnych danych, a warunki brzegowe symulacji skupiały się na wysokim poziomie związków między pozycjami (sytuacjach z punktu widzenia psychologii raczej idealnych i niezbyt często obserwowanych) oraz c) główny nacisk artykułu został położony na sytuację, w której poprawny model jest oceniany jako niepoprawny, a nie odwrotnie – niepoprawny (niepoprawnie wyspecyfikowany) jako poprawny.

Istnieją także inne publikacje o zbliżonej tematyce, to znaczy nieprawidłowej specyfikacji modeli, niemniej jednak nie skupiają się one na praktycznych zagadnieniach (czyli na przykład próbie określenia prostych i pewnych kryteriów decyzyjnych, które mogą bez wątpliwości być podstawą rozstrzygnięcia, czy struktura operacjonalizowanego zjawiska jest jedno- czy wielowymiarowa), a jedynie eksplorują samą kwestię zagadnień ewentualnych konsekwencji błędnej specyfikacji lub różnych możliwości potencjalnego wykrywania błędów specyfikacyjnych bez wskazania rozwiązań danych problemów czy nawet określania, co konkretnie jest przyczyną niepoprawnej specyfikacji (na przykład: Dennis, Ponciano, Taper i Lele, 2019; Rhemtulla, 2016; Saris, et al., 2009). Niektóre z tych artykułów pośrednio poruszają problemy, które były obiektem zainteresowania niniejszego projektu i prezentują zbliżone wyniki do tu zaprezentowanych – to znaczy wskazują na różne wady kryteriów dopasowania oraz problematyczność w ich wykorzystaniu w kontekście rozpoznawania nieprawidłowej specyfikacji modelu, podważając w mniejszym lub większym stopniu całkowitą poprawność wykrywania błędów specyfikacji z użyciem popularnych metod (na

przykład: Wang, Xu, Wang, Tan i Chen, 2020; Preacher i Merkle, 2012; Jarvis, MacKenzie i Podsakoff, 2012; Yang i Green, 2010; Fan i Sivo, 2007, 2005). Ponownie, nie eksplorują one jednak kwestii przyczyn (nie)poprawnej specyfikacji, i żadne z wymienionych badań do tej pory nie sprawdzało, czy na tych samych danych można dopasowywać modele o różnej liczbie założonych czynników latentnych w stopniu pozwalającym na ich akceptację (w zdecydowanej większości przytoczone badania symulacyjne po prostu sprawdzają, czy dany próg dopasowania pozwolił na akceptację modelu lub nie). Wygląda zatem na to, że w niniejszej pracy po raz pierwszy a) dokonano eksploracji jednej konkretnej przyczyny nieprawidłowej specyfikacji pod kątem poprawności oceny tego, na ile miary dopasowania mogą wspomagać badacza w decyzji co do wymiarowości jego narzędzia w analizie CFA; b) wykorzystano w badaniach wszystkie dostępne (popularne) miary dopasowania i porównano różne modele oparte o te same dane między sobą, szukając różnic między ich miarami dopasowania; c) zastosowano szeroki zakres warunków wejściowych symulacji, aby określić także bardziej realistyczne warianty możliwych modeli (nie tylko idealne); d) potwierdzono wyniki symulacji analizami przeprowadzonymi na realnych wynikach empirycznych; e) zaproponowano wykorzystanie w określaniu nieprawidłowej specyfikacji modeli pod kątem liczby wymiarów narzędzia psychometrycznego podejście oparte o indeksy modyfikacji.

6.2 Rekomendacje

W świetle przedstawionych wyników można sformułować następujące rekomendacje:

(1) Nie należy traktować żadnych globalnych miar dopasowania modelu jako determinujących, czy rzeczywista operacjonalizowana zmienna psychologiczna jest jedno-, czy wielowymiarowa.

(2) Do określania poprawności modelu w kontekście odróżnienia jedno- od wielowymiarowych struktur można wykorzystać niektóre miary oparte o indeksy modyfikacji.

(3) Rekomenduje się zastosowanie miary *nm* do porównania alternatywnych struktur tego samego narzędzia (wyższa wartość określa lepszy model). Można także wykorzystać miary *EPC* (wartość większa niż 0,010 może wskazywać błędną liczbę czynników w modelu), choć względem tej drugiej miary należy zachować pewną ostrożność we wnioskowaniu.

Dodatkowo, można zarekomendować, aby w przypadku zbliżonego (lub akceptowalnego w obu przypadkach) dopasowania modeli jedno- i wielowymiarowych (na tym

samym zbiorze danych) przyjmować rozwiązanie jednowymiarowe lub dodatkowo posłużyć się analizami o charakterze eksploracyjnym.

Zaproponowany przez Saris, Satorra i Van der Veld (2009) i zaimplementowany w R przez Uanhoro (2017) algorytm oparty o wykorzystanie indeksów modyfikacji wydaje się być obecnie mało popularnym narzędziem, choć jest to narzędzie relatywnie łatwe do wykorzystania i mogłoby być dla psychologów nowym standardem w podejściu do identyfikowania jednowymiarowych struktur. Możliwe, że problemem jest obecnie różna dostępność tej metody w różnych pakietach statystycznych (pakiety do równań strukturalnych w R zdają się być mniej popularne w psychologii niż wizualne oprogramowanie typu AMOS). Pozostawiono zatem w materiałach dodatkowych pod linkiem https://osf.io/y36hz prosty skrypt pozwalający na określenie miary indeksów dopasowania, o których tu mowa.

6.3 Ograniczenia badań własnych oraz kierunki dalszych badań

Zaprezentowane wyniki wskazują wyraźnie, że w przypadku, gdy struktura danego kwestionariusza jest jednowymiarowa, to z użyciem wszystkich wymienionych miar dopasowania i większości wskaźników opartych o indeksy modyfikacji, można ją bez większego problemu opisać jako wielowymiarową, biorąc pod uwagę nawet współczesne, coraz bardziej konserwatywne standardy publikacyjne. Z praktycznego punktu widzenia, badacz porównujący dwa alternatywne modele na podstawie wskaźników dopasowania ma jednakową szansę na uznanie każdego z nich za lepszy, a różnice między tymi modelami mogą być po prostu losowe lub pomijalne. Należy przy tym podkreślić, że mówimy tu o sytuacji idealnej, w której wiadomo, że jeden czynnik jest poprawny (jest silnym i jedynym czynnikiem w obrębie danych), a podczas badań psychologicznych często struktura operacjonalizowanego konstruktu jest nieznana lub niepewna (związki pomiędzy poszczególnymi itemami są słabsze, a wymiary zanieczyszczone na przykład niespójnymi itemami). Wprawdzie badaczy zachęca się czasem do wyboru prostszych struktur (Brauer i Ranger, 2021), ale przecież ta prostsza struktura może w ogóle nie być przez badacza uwzględniana w obrębie jego teorii.

Istotnym ograniczeniem niniejszej pracy jest skupienie się wyłącznie na klasycznej konfirmacyjnej analizie czynnikowej (CFA) jako metodzie oceny poprawności struktury zmiennej latentnej. Choć CFA stanowi obecnie jedno z najczęściej stosowanych podejść w psychometrii i dostarcza dobrze opracowanego aparatu statystycznego oraz interpretacyjnego, nie jest jedyną metodą służącą do oceny wymiarowości konstruktów. W literaturze coraz częściej wskazuje się na ograniczenia klasycznej CFA – w tym na przykład sztywność założeń dotyczących niezależności błędów pomiędzy itemami czy konieczność

z góry ustalonej struktury pomiarowej (na przykład: Pan, Ip, i Dubé, 2017). W związku z tym w ostatnich latach coraz większe zainteresowanie budzą alternatywne podejścia, takie jak modele bifactorowe (bifactor CFA, na przykład: Murrayi in., 2019), eksploracyjno-konfirmacyjne modele czynnikowe (*ESEM, Exploratory Structural Equation Modeling, Eksploracyjna Analiza Równań Strukturalnych*, na przykład: Morin, Myers, i Lee, 2020), czy też ujęcia bayesowskie (Lee, 1981) i modele oparte o teorię odpowiedzi na itemy (*IRT, Item Response Theory*, Reise,i in. 1993).

Modele bifactorowe pozwalają na jednoczesne modelowanie wspólnego czynnika ogólnego oraz czynników specyficznych, co bywa szczególnie przydatne przy analizie konstrukcji rzekomo jednowymiarowych, które mogą zawierać ukryte czynniki drugorzędne (Cheni in., 2012). Podejście ESEM, będące rozwinięciem klasycznego CFA, łączy w sobie zalety analiz eksploracyjnych i konfirmacyjnych, pozwalając na większą elastyczność w modelowaniu danych bez rezygnowania z możliwości testowania hipotez teoretycznych (Marshi in., 2020). Z kolei modele IRT (szczególnie modele wielowymiarowe) umożliwiają ocenę trafności pozycji oraz struktury wymiarowej konstruktu z perspektywy właściwości poszczególnych itemów, co może dostarczyć bardziej szczegółowych informacji diagnostycznych (Levy & Mislevy, 2017; Raju, Laffitte i Byrne, 2002). Zastosowanie tych alternatywnych podejść w przyszłych badaniach mogłoby znacząco poszerzyć rozumienie trafności strukturalnej narzędzi pomiarowych oraz pozwolić na ocenę odporności klasycznych wskaźników CFA na różne założenia i warunki danych. Co więcej, porównanie wyników uzyskiwanych przy pomocy klasycznego CFA z wynikami pochodzącymi z modeli bifactorowych czy ESEM mogłoby dostarczyć informacji, które z podejść są bardziej adekwatne przy decyzji o liczbie wymiarów w konstrukcie. Tym samym możliwe byłoby zbudowanie szerszych rekomendacji dla praktyków psychometrii, wykraczających poza ograniczenia tradycyjnych technik analitycznych.

W niniejszej pracy zastosowano wyłącznie klasyczną metodę estymacji maksymalnej wiarygodności (ML, *Maximum Likelihood*), pomijając inne dostępne metody, takie jak uogólniona najmniejszych kwadratów (GLS, *Generalized Least Squares*), asymptotycznie niezależna (ADF, *Asymptotic Distribution-Free*) czy estymacja robocza (*WLS, Weighted Least Squares - Ważona Metoda Najmniejszych Kwadratów*, lub *DWLS Diagonally Weighted Least Squares, Diagonalna Ważona Metoda Najmniejszych Kwadratów*). Wybór estymacji metodą ML jest zgodny z dominującą praktyką badawczą w symulacjach i zapewnia dobrą efektywność estymatorów w warunkach zgodności danych z założeniami rozkładu normalnego (Brown, 2015), jednak skuteczność i stabilność poszczególnych miar dopasowania może

zależeć od wybranej metody estymacji (Sharmai in., 2005; DiStefano i Morgan, 2014). Przykładowo, metody GLS i ADF są rekomendowane w sytuacjach, gdy dane odbiegają od rozkładu normalnego lub mają nietypowe właściwości statystyczne (Kline, 2015). Ponadto niektóre współczynniki dopasowania (np. RMSEA czy CFI) mogą przyjmować inne wartości w zależności od zastosowanego algorytmu estymacyjnego, co ogranicza możliwość generalizacji uzyskanych w niniejszej pracy wyników do sytuacji, w których by zastosowano inne możliwe metody estymacji. W przyszłych badaniach warto zatem przeprowadzić porównania działania wskaźników dopasowania w zależności od metody estymacji, szczególnie w kontekście danych o różnym stopniu normalności, z uwzględnieniem różnych wielkości prób czy liczby itemów.

Kolejnym ograniczeniem badania jest naturalnie zakres wybranych do porównania i opracowania typów modeli CFA. W obrębie niniejszej pracy nie weryfikowano szczegółowych różnic między modelami dwu- i więcej czynnikowymi, ponieważ praca skupiała się głównie na rozróżnieniu modeli jedno- i (zbiorczo wszystkich) wielowymiarowych. Dodatkowo nie były w obrębie prowadzonych badań weryfikowane symulowane struktury dwu- i więcej wymiarowe, a więc nie testowano czy zaprezentowane tu wyniki można ekstrapolować na określenie prawidłowości wyspecyfikowania modeli wielowymiarowych. Innymi słowy analizowana była sytuacja w której dla struktury jednowymiarowej badacz zakłada strukturę wielowymiarową, ale nie odwrotna, czyli gdy badacz dla struktury wielowymiarowej błędnie zakłada strukturę jednowymiarową. Zatem przyszłe badania mogłyby objąć przypadek odwrotny – testowanie trafności modeli jednowymiarowych wobec danych o strukturze wielowymiarowej. Finalnie – można dla różnych struktur wielowymiarowych sprawdzać możliwe dopasowanie jako struktur również wielowymiarowych, ale o innej liczbie czynników latentnych – na przykład dla struktur dwuwymiarowych sprawdzać możliwość dopasowania ich jako struktur trzywymiarowych i odwrotnie, oczywiście uwzględniając wszystkie wymienione w badaniu kontrolowane warunki symulacji oraz wymienione w dyskusji potencjalne dodatkowe moderatory wyników. W przypadku struktur wielowymiarowych należałoby także jako dodatkową rozważyć kwestię błędnej alokacji itemów, czyli sytuacji, gdy badacz w modelu CFA błędnie przypisuje niektóre itemy do nieodpowiednich czynników latentnych (na przykład: item faktycznie mierzy czynnik A, ale w modelu badacz przypisuje go do czynnika B). To wymagałoby nie tylko określenia, czy w strukturze itemy są błędnie alokowane, ale także tego, jaka liczba (jaki procent) błędnej alokacji itemów ma znaczenie dla wyników analizy.

W dalszych badaniach należałoby zweryfikować różne poziomy zanieczyszczenia zmiennych latentnych (lub różne metody zanieczyszczania danych), a także poziom korelacji pomiędzy wymiarami struktur wielowymiarowych (w tym także w różnych kombinacjach proporcji korelacji pomiędzy więcej niż dwoma czynnikami). W każdej z możliwych dalszych badań należałoby także zebrać odpowiednie dane empiryczne o podobnym poziomie zanieczyszczenia lub związków między wymiarami oraz wyraźnie odpowiednio wielowymiarowej strukturze. Jest to zatem kolejny potencjalny zakres kierunków przyszłych badań w tym zakresie.

Mimo że w badaniu uwzględniono szeroki zakres wariantów danych zarówno symulacyjnych, jak i empirycznych, należy zwrócić uwagę na pewne potencjalne ograniczenie natury metodologicznej. Zarówno dane empiryczne, jak i wygenerowane dane symulacyjne, przyjmowały założenie „czystych" odpowiedzi – to znaczy odpowiedzi niezaburzonych wpływami systematycznymi, np. tendencją do udzielania społecznie aprobowanych odpowiedzi, efektami kolejności itemów, czy błędami wynikającymi z niejednoznaczności językowej. W praktyce psychologicznej występowanie takich zakłóceń jest powszechne i może wpływać nie tylko na strukturę danych, ale również na zachowanie miar dopasowania. W przyszłych badaniach warto więc rozważyć wprowadzenie do modeli symulacyjnych elementów błędu systematycznego (np. poprzez dodanie do niektórych itemów efektów „biasu" lub korelacji reszt), aby sprawdzić, na ile klasyczne i lokalne miary dopasowania są wrażliwe na tego typu zniekształcenia. Taka procedura mogłaby jeszcze bardziej zbliżyć warunki analizy do realiów pracy z danymi psychologicznymi.

Oczywiście niniejsze badania symulacyjne i empiryczne nie wyczerpują wszystkich możliwości wejściowych warunków badawczych. Można przecież tworzyć modele o większym poziomie skomplikowania, różnym poziomie zanieczyszczenia czynników, testować poprawnie dopasowane modele wielowymiarowe, sprawdzać bardzo duże i bardzo małe próby i tak dalej. Można także szukać innych dróg określania poprawności modelu, mniej lub bardziej zaawansowanych. Pojawia się pytanie, czy można rozstrzygnąć w obrębie stosowanych powszechnie metod analizy liczby zmiennych latentnych, czy mała różnica między punktami wykresów osypiska jest już wielowymiarowością, czy zanieczyszczeniem. Nie jest jasne, jak duża jest różnica faktyczna między modelami jedno- i wielowymiarowymi jeśli chodzi o wskaźniki określające ich wymiarowość. Ostatecznie rozstrzygnięcia wymaga kwestia, czy automatyczne metody redukcji i ustalania czynników zamiast ustalać, ile czynników ma jakaś struktura, po prostu preferują jej określone właściwości, nie tyle świadczące o wymiarowości, co określające coś, co z wymiarowością ma tylko pośredni związek.

Wyniki, które tu są zaprezentowane, można w zasadzie potraktować jako początek poszukiwań jednoznacznych kryteriów pozwalających uniknąć błędu polegającego na potwierdzaniu modeli wielowymiarowych, których struktura w istocie jest jednowymiarowa. Z pewnością przyszłe badania w tym zakresie będą potrzebne, zwłaszcza w kontekście ogromnej różnorodności konfirmowanych modeli odpowiadających złożonym zjawiskom psychologicznym. Finalnie, można mieć nadzieję, że wyniki tu zaprezentowane i oparte na nich podejście może poprawić jakość tworzonych narzędzi badawczych, oraz pozwoli poprawnie identyfikować liczbę wymiarów zmiennych latentnych i przyczynić się do rozwoju psychometrii.

BIBLIOGRAFIA

Aguirre-Urreta, M. I., Marakas, G. M. i Ellis, M. E. (2013). Measurement of composite reliability in research using partial least squares: Some issues and an alternative approach. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, *44*(4), 11-43. https://doi.org/10.1145/2544415.2544417

Alin, A. (2010). Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(3), 370-374. https://doi.org/10.1002/wics.84

André, F., Munck, I., Håkansson, A. i Claesdotter-Knutsson, E. (2022). Game Addiction Scale for Adolescents—Psychometric Analyses of Gaming Behavior, Gender Differences and ADHD. *Frontiers in Psychiatry*, *13*, 791254. https://doi.org/10.3389/fpsyt.2022.791254

Auerswald, M. i Moshagen, M. (2019). How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions. *Psychological Methods*, *24*(4), 468. https://doi.org/10.1037/met0000200

Ayanso, A. i Mokaya, B. (2013). Efficiency evaluation in search advertising. *Decision Sciences*, *44*(5), 877-913. https://doi.org/10.1111/deci.12038

Baldwin, S. A. (2019). *Psychological statistics and psychometrics using Stata* (p. 454). College Station, TX: Stata Press.

Bannigan, K. i Watson, R. (2009). Reliability and validity in a nutshell. *Journal of Clinical Nursing*, *18*(23), 3237-3243. https://doi.org/10.1111/j.1365-2702.2009.02939.x

Barendse, M. T., Oort, F. J. i Timmerman, M. E. (2015). Using exploratory factor analysis to determine the dimensionality of discrete responses. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(1), 87-101. https://doi.org/10.1080/10705511.2014.934850

Bartholomew, D. J., Knott, M. i Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. John Wiley i Sons. https://doi.org/10.1002/9781119970583

Bhattacherjee, A. (2012). *Social science research: Principles, methods, and practices*. USA.

Blunch, N. J. (2012). Introduction to structural equation modeling using IBM SPSS statistics and AMOS. https://doi.org/10.4135/9781526402257

Bollen, K. A. (1990). Overall fit in covariance structure models: Two types of sample size effects. *Psychological Bulletin, 107*(2), 256. https://doi.org/10.1037/0033-2909.107.2.256

Bollen, K. A. i Noble, M. D. (2011). Structural equation models and the quantification of behavior. *Proceedings of the National Academy of Sciences*, *108*(supplement_3), 15639-15646. https://doi.org/10.1073/pnas.1010661108

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425-440. https://doi.org/10.1007/s11336-006-1447-6

Bowen, N. K. i Guo, S. (2011). *Structural equation modeling*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195367621.001.0001

Braeken, J. i Van Assen, M. A. (2017). An empirical Kaiser criterion. *Psychological Methods*, *22*(3), 450. https://doi.org/10.1037/met0000074

Brauer, K. i Ranger, J. (2021). Determining the Number of Factors in Exploratory Factor Analysis–The Performance of Tests of Relative Model Fit. *Psychological Test and Assessment Modeling*, *63*(1), 45-64.

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research* (2nd ed.). The Guilford Press.

Brunner, M. i Süβ, H. M. (2005). Analyzing the reliability of multidimensional measures: An example from intelligence research. *Educational and Psychological Measurement*, *65*(2), 227-240. https://doi.org/10.1177/0013164404268669

Byrne, B. M. (2005). Factor analytic models: Viewing the structure of an assessment instrument from three perspectives. *Journal of Personality Assessment*, *85*(1), 17-32. https://doi.org/10.1207/s15327752jpa8501_02

Cerreia-Vioglio, S., Hansen, L. P., Maccheroni, F. i Marinacci, M. (2020). Making decisions under model misspecification. *arXiv preprint arXiv:2008.01071*. https://doi.org/10.2139/ssrn.3666424

Chakrabarti, A. i Ghosh, J. K. (2011). AIC, BIC and recent advances in model selection. *Philosophy of Statistics*, 583-605. https://doi.org/10.1016/B978-0-444-51862-0.50018-6

Chalmers, R. P. i Adkins, M. C. (2020). Writing effective and reliable Monte Carlo simulations with the SimDesign package. *The Quantitative Methods for Psychology*, *16*(4), 248-280. https://doi.org/10.20982/tqmp.16.4.p248

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J. P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor

model to other approaches. *Journal of Personality*, *80*(1), 219-251. https://doi.org/10.1111/j.1467-6494.2011.00739.x

Chen, G., Gully, S. M. i Eden, D. (2004). General self-efficacy and self-esteem: Toward theoretical and empirical distinction between correlated self-evaluations. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, *25*(3), 375-395. https://doi.org/10.1002/job.251

Cheng, C., Lay, K. L., Hsu, Y. F. i Tsai, Y. M. (2021). Can Likert scales predict choices? Testing the congruence between using Likert scale and comparative judgment on measuring attribution. *Methods in Psychology*, *5*, 100081. https://doi.org/10.1016/j.metip.2021.100081

Cheung, G. W., Cooper-Thomas, H. D., Lau, R. S. i Wang, L. C. (2024). Reporting reliability, convergent and discriminant validity with structural equation modeling: A review and best-practice recommendations. *Asia Pacific Journal of Management*, *41*(2), 745-783. https://doi.org/10.1007/s10490-023-09871-y

Cheung, G. W. i Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling*, *9*(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5

Civelek, M. E. (2018). *Essentials of structural equation modeling*. Lulu.com. https://doi.org/10.13014/K2SJ1HR5

Clayton, M. F. i Pett, M. A. (2008). AMOS versus LISREL: One data set, two analyses. *Nursing research*, *57*(4), 283-292. https://doi.org/10.1097/01.NNR.0000313487.64412.be

Collier, J. (2020). *Applied structural equation modeling using AMOS: Basic to advanced techniques*. Routledge. https://doi.org/10.4324/9781003018414

Collins, L. M. i Lanza, S. T. (2009). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). John Wiley i Sons. https://doi.org/10.1002/9780470567333

Cooper, C. (2023). *An Introduction to Psychometrics and Psychological Assessment: Using, Interpreting and Developing Tests*. Routledge. https://doi.org/10.4324/9781003240181

Cudeck, R. i MacCallum, R. C. (Eds.). (2007). *Factor analysis at 100: Historical developments and future directions*. Routledge. https://doi.org/10.4324/9780203936764

Cunningham, J. P. i Ghahramani, Z. (2015). Linear dimensionality reduction: Survey, insights, and generalizations. *The Journal of Machine Learning Research*, *16*(1), 2859-2900.

Dalal, D. K. (2023). (Multi) Collinearity in Behavioral Sciences Research. In *Oxford Research Encyclopedia of Business and Management*. https://doi.org/10.1093/acrefore/9780190224851.013.410

Daoud, J. I. (2017). Multicollinearity and regression analysis. In *Journal of Physics: Conference Series* (Vol. 949, No. 1, p. 012009). IOP Publishing. https://doi.org/10.1088/1742-6596/949/1/012009

de la Rubia, J. M. (2022). Note on Rank-Biserial Correlation When There Are Ties. *Open Journal of Statistics*, *12*(5), 597-622. https://doi.org/10.4236/ojs.2022.125036

Dennis, B., Ponciano, J. M., Taper, M. L. i Lele, S. R. (2019). Errors in statistical inference under model misspecification: Evidence, hypothesis testing, and AIC. *Frontiers in Ecology and Evolution*, *7*, 372. https://doi.org/10.3389/fevo.2019.00372

Dimov, I. T. (2008). *Monte Carlo methods for applied scientists*. World Scientific. https://doi.org/10.1142/9789812779892

Dunn, W. L. i Shultis, J. K. (2011). *Exploring monte carlo methods*. Elsevier. https://doi.org/10.1016/B978-0-444-51575-9.00007-5

El-Den, S., Schneider, C., Mirzaei, A. i Carter, S. (2020). How to measure a latent construct: Psychometric principles for the development and validation of measurement instruments. *International Journal of Pharmacy Practice*, *28*(4), 326-336. https://doi.org/10.1111/ijpp.12600

Evermann, J. i Tate, M. (2010). Testing models or fitting models? Identifying model misspecification in PLS. ICIS 2010 Proceedings. 21

Fan, X. i Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling*, *12*(3), 343-367. https://doi.org/10.1207/s15328007sem1203_1

Fan, X. i Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research*, *42*(3), 509-529. https://doi.org/10.1080/00273170701382864

Fialkowski, A. i Tiwari, H. (2019). SimCorrMix: Simulation of Correlated Data with Multiple Variable Types Including Continuous and Count Mixture Distributions. *R J.*, *11*(1), 250. https://doi.org/10.32614/RJ-2019-022

Finch, H., French, B. F. i Immekus, J. C. (2016). *Applied psychometrics using SPSS and AMOS*. IAP.

Finch, W. H. (2020). Using fit statistic differences to determine the optimal number of factors to retain in an exploratory factor analysis. *Educational and Psychological Measurement*, *80*(2), 217-241. https://doi.org/10.1177/0013164419865769

Fisk, C. L., Harring, J. R., Shen, Z., Leite, W., Suen, K. Y. i Marcoulides, K. M. (2023). Using simulated annealing to investigate sensitivity of SEM to external model misspecification. *Educational and Psychological Measurement*, *83*(1), 73-92. https://doi.org/10.1177/00131644211073121

Fuoli, M. (2022). Structural Equation Modeling in R: a practical introduction for linguists. *Data Analytics in Cognitive Linguistics: Methods and Insights*, *41*, 75. https://doi.org/10.1515/9783110687279-004

Furr, R. M. (2021). *Psychometrics: an introduction*. SAGE publications.

Garrido, L. E., Abad, F. J. i Ponsoda, V. (2016). Are fit indices really fit to estimate the number of factors with categorical variables? Some cautionary findings via Monte Carlo simulation. *Psychological Methods*, *21*(1), 93. https://doi.org/10.1037/met0000064

Garson, G. D. (2022). *Factor analysis and dimension reduction in R: A social scientist's toolkit*. Routledge. https://doi.org/10.4324/9781003279693

Gazdowska, Z., Ryszkiewicz, M. i Parzelski, D. (2015). Polska adaptacja Kwestionariusza Źródeł Pewności Siebie w Sporcie (SSCQ-PL)–geneza teoretyczna oraz badanie pilotażowe właściwości psychometrycznych. *Journal of Education, Health and Sport*, *5*(10), 207-220.

Goldfeld, K., Wujciak-Jens, J. (2020). simstudy: Illuminating research methods through data generation. *Journal of Open Source Software,* *5*(54), 2763. https://doi.org/10.21105/joss.02763

Greenacre, M., Groenen, P. J., Hastie, T., d'Enza, A. I., Markos, A. i Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, *2*(1), 100. https://doi.org/10.1038/s43586-022-00184-w

Greiff, S. i Heene, M. (2017). Why psychological assessment needs to start worrying about model fit. *European Journal of Psychological Assessment*. *33*(5), 313-317 https://doi.org/10.1027/1015-5759/a000450

Gulliksen, H. (2013). *Theory of mental tests*. Routledge. https://doi.org/10.4324/9780203052150

Gustafsson, J. E. i Åberg-Bengtsson, L. (2010). *Unidimensionality and interpretability of psychological instruments*. https://doi.org/10.1037/12074-005

Hancock, G. R. i Mueller, R. O. (Eds.). (2013). *Structural equation modeling: A second course*. Iap.

Harding, B., Tremblay, C. i Cousineau, D. (2014). Standard errors: A review and evaluation of standard error estimators using Monte Carlo simulations. *The Quantitative Methods for Psychology*, *10*(2), 107-123. https://doi.org/10.20982/tqmp.10.2.p107

Havigerová, J. M., Dosedlová, J. i Burešová, I. (2018). One health behavior or many health-related behaviors?. *Psychology Research and Behavior Management*, 23-30. https://doi.org/10.2147/PRBM.S173692

Hayashi, K., Bentler, P. M. i Yuan, K. H. (2007). 13 Structural Equation Modeling. *Handbook of statistics*, *27*, 395-428. https://doi.org/10.1016/S0169-7161(07)27013-0

Hayes, N. (2021). *Doing Psychological Research*, 2e. McGraw-Hill Education (UK).

Headey, D. (2008). The principal components of growth in the less developed countries. *Kyklos*, *61*(4), 568-598. https://doi.org/10.1111/j.1467-6435.2008.00418.x

Henley, S. S., Golden, R. M. i Kashner, T. M. (2020). Statistical modeling methods: challenges and strategies. *Biostatistics i Epidemiology*, *4*(1), 105-139. https://doi.org/10.1080/24709360.2019.1618653

Henseler, J. (2012). Why generalized structured component analysis is not universally preferable to structural equation modeling. *Journal of the Academy of Marketing Science*, *40*, 402-413. https://doi.org/10.1007/s11747-011-0298-6

Hill, B. D. (2011). *The sequential Kaiser-Meyer-Olkin procedure as an alternative for determining the number of factors in common-factor analysis: A Monte Carlo simulation*. Oklahoma State University.

Hooper, D., Coughlan, J., Mullen, M. R., Mullen, J., Hooper, D., Coughlan, J. i Mullen, M. R. (2008). Structural equation modelling: guidelines for determining model fit structural equation modelling: guidelines for determining model fit. Dublin Institute of Technology ARROW@ DIT, 6 (1), 53–60.

Hoyle, R. H. (Ed.). (2012). *Handbook of structural equation modeling*. Guilford press.

Iacobucci, D. (2010). Structural equations modeling: Fit indices, sample size, and advanced topics. *Journal of Consumer Psychology*, *20*(1), 90-98. https://doi.org/10.1016/j.jcps.2009.09.003

Ilska, M. i Przybyła-Basista, H. (2020). Psychometric evaluation of the Polish version of the Support in Intimate Relationships Rating Scale–Revised (SIRRS-R). *Current Issues in Personality Psychology*, *8*(2), 139-153. https://doi.org/10.5114/cipp.2020.94261

Jackson, D. L., Gillaspy Jr, J. A. i Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: an overview and some recommendations. *Psychological Methods*, *14*(1), 6. https://doi.org/10.1037/a0014694

Jahanian, R. i Poornaghi, Z. (2012). The relationship between state-trait anxiety and students' sense of social self-efficacy. *World Applied Sciences Journal*, *20*(3), 395-400.

Jarvis, C. B., MacKenzie, S. B. i Podsakoff, P. M. (2012). The negative consequences of measurement model misspecification: A response to Aguirre-Urreta and Marakas. *MIS quarterly*, *36*(1), 139-146. https://doi.org/10.2307/41410410

Jöreskog, K. G. i Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Scientific software international.

Juczyński, Z. (2000). Poczucie własnej skuteczności – teoria i pomiar. *Acta Univ. Lodz. Folia Psychol. 2000, 4*, 11–24.

Kalnins, A. (2018). Multicollinearity: How common factors cause Type 1 errors in multivariate regression. *Strategic Management Journal*, *39*(8), 2362-2385. https://doi.org/10.1002/smj.2783

Kalnins, A. (2022). When does multicollinearity bias coefficients and cause type 1 errors? A reconciliation of Lindner, Puck, and Verbeke (2020) with Kalnins (2018). *Journal of international business studies*, *53*(7), 1536-1548. https://doi.org/10.1057/s41267-022-00531-9

Kaźmierczak, M. i Plopa, M. (2008). *Kwestionariusz komunikacji małżeńskiej: podręcznik*. Vizja Press i IT.

Kenny, D. (2015). Measuring model fit. Retrieved November, 2023, from http://davidakenny.net/cm/fit.htm

Khademi, A. (2017). Multivariate Analysis with LISREL. *Journal of Statistical Software*, *78*, 1-3. https://doi.org/10.18637/jss.v078.b01

Khattree, R. i Naik, D. N. (2000). Multivariate data reduction and discrimination. *SAS Institute, Cary, North Carolina*.

Kim, E. S. (2011). *Testing measurement invariance using MIMIC: Likelihood ratio test and modification indices with a critical value adjustment*. Texas A&M University. https://doi.org/10.1177/0013164411427395

Kim, J. H. (2019). Multicollinearity and misleading statistical results. *Korean journal of anesthesiology*, *72*(6), 558-569. https://doi.org/10.4097/kja.19087

Kline, P. (2015). *A handbook of test construction (psychology revivals): introduction to psychometric design*. Routledge. https://doi.org/10.4324/9781315695990

Kline, R. B. (2010). Promise and pitfalls of structural equation modeling in gifted research. In B. Thompson i R. F. Subotnik (Eds.), *Methodologies for conducting research on giftedness* (pp. 147–169). American Psychological Association. https://doi.org/10.1037/12079-007

Kline, R. B. (2023). *Principles and practice of structural equation modeling*. Guilford publications.

Koliopanos, G., Ojeda, F. i Ziegler, A. (2023). A Simple-to-Use R Package for Mimicking Study Data by Simulations. *Methods of Information in Medicine*, *62*(03/04), 119-129. https://doi.org/10.1055/a-2048-7692

Komlos, J. (2019). Multicollinearity in the presence of errors-in-variables can increase the probability of type-I error. https://doi.org/10.2139/ssrn.3338813

Konarski, R. (2009). *Modele równań strukturalnych: teoria i praktyka*. Wydawnictwo Naukowe PWN.

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological methods i research*, *33*(2), 188-229. https://doi.org/10.1177/0049124103262065

Kyriazos, T. A. (2018). Applied psychometrics: sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. *Psychology*, *9*(08), 2207. https://doi.org/10.4236/psych.2018.98126

Kyriazos, T. i Poga-Kyriazou, M. (2023). Applied Psychometrics: Estimator Considerations in Commonly Encountered Conditions in CFA, SEM, and EFA Practice. *Psychology*, *14*(5), 799-828. https://doi.org/10.4236/psych.2023.145043

Lamers, S. M., Westerhof, G. J., Bohlmeijer, E. T., ten Klooster, P. M. i Keyes, C. L. (2011). Evaluating the psychometric properties of the mental health continuum-short form (MHC-SF). *Journal of Clinical Psychology*, *67*(1), 99-110. https://doi.org/10.1002/jclp.20741

Lance, C. E., Butts, M. M. i Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say?. *Organizational Research Methods*, *9*(2), 202-220. https://doi.org/10.1177/1094428105284919

Ledesma, R. D., Valero-Mora, P. i Macbeth, G. (2015). The scree test and the number of factors: a dynamic graphics approach. *The Spanish Journal of Psychology*, *18*, E11. https://doi.org/10.1017/sjp.2015.13

Lee, M. D. (2001). Determining the dimensionality of multidimensional scaling representations for cognitive modeling. *Journal of Mathematical Psychology*, *45*(1), 149-166. https://doi.org/10.1006/jmps.1999.1300

Lee, S. Y. (1981). A Bayesian approach to confirmatory factor analysis. *Psychometrika*, *46*(2), 153-160. https://doi.org/10.1007/BF02293896

Levy, R., & Mislevy, R. J. (2017). *Bayesian psychometric modeling*. Chapman and Hall/CRC. https://doi.org/10.1201/9781315374604

Lightsey Jr, O. R., Burke, M., Ervin, A., Henderson, D. i Yee, C. (2006). Generalized self-efficacy, self-esteem, and negative affect. *Canadian Journal of Behavioural Science/Revue Canadienne des Sciences du Comportement*, *38*(1), 72. https://doi.org/10.1037/h0087272

Lima-Rodríguez, J. S., Lima-Serrano, M. i Domínguez-Sánchez, I. (2015). Psychometric properties of an instrument to measure family disease management. *International Journal of Clinical and Health Psychology*, *15*(3), 253-264. https://doi.org/10.1016/j.ijchp.2015.05.002

Lin, L. C., Huang, P. H. i Weng, L. J. (2017). Selecting path models in SEM: A comparison of model selection criteria. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(6), 855-869. https://doi.org/10.1080/10705511.2017.1363652

Loewenthal, K. M. i Lewis, C. A. (2020). *An introduction to psychological tests and scales*. Routledge.

Lorenzo-Seva, U., Timmerman, M. E. i Kiers, H. A. (2011). The Hull method for selecting the number of common factors. *Multivariate Behavioral Research*, *46*(2), 340-364. https://doi.org/10.1080/00273171.2011.564527

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour Research and Therapy*, *33*(3), 335-343. https://doi.org/10.1016/0005-7967(94)00075-u

Łaguna, M., Lachowicz-Tabaczek, K. i Dzwonkowska, I. (2007). Skala samooceny SES Morrisa Rosenberga–polska adaptacja metody. *Psychologia Społeczna*, *2*(4), 164-176.

MacFarland, T. W., Yates, J. M., MacFarland, T. W. i Yates, J. M. (2016). Mann–whitney u test. *Introduction to nonparametric statistics for the biological sciences using R*, 103-132. https://doi.org/10.1007/978-3-319-30634-6_4

Mair, P. (2018). *Modern psychometrics with R*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-93177-7

Makara-Studzińska, M., Załuski, M., Adamczyk, K. i Tyburski, E. (2022). Polska wersja Skali Depresji Lęku i Stresu (DASS-42)–adaptacja i normalizacja. *Psychiatria Polska*.

Marcoulides, G. A. i Schumacker, R. E. (Eds.). (2013). *Advanced structural equation modeling: Issues and techniques*. Psychology Press. https://doi.org/10.4324/9781315827414

Marella, M., Pesudovs, K., Keeffe, J. E., O'Connor, P. M., Rees, G. i Lamoureux, E. L. (2010). The psychometric validity of the NEI VFQ-25 for use in a low-vision population. *Investigative Ophthalmology i Visual Science*, *51*(6), 2878-2884. https://doi.org/10.1167/iovs.09-4494

Margolis, S., Schwitzgebel, E., Ozer, D. J. i Lyubomirsky, S. (2019). A new measure of life satisfaction: The Riverside Life Satisfaction Scale. *Journal of Personality Assessment*, *101*(6), 621-630. https://doi.org/10.1080/00223891.2018.1464457

Markland, D. (2007). The golden rule is that there are no golden rules: A commentary on Paul Barrett's recommendations for reporting model fit in structural equation modelling. *Personality and Individual Differences*, *42*(5), 851-858. https://doi.org/10.1016/j.paid.2006.09.023

Marsh, H. W., Guo, J., Dicke, T., Parker, P. D., & Craven, R. G. (2020). Confirmatory factor analysis (CFA), exploratory structural equation modeling (ESEM), and set-ESEM: optimal balance between goodness of fit and parsimony. *Multivariate behavioral research*, *55*(1), 102-119. https://doi.org/10.1080/00273171.2019.1602503

Maydeu-Olivares, A. (2017). Assessing the size of model misfit in structural equation models. *Psychometrika*, *82*(3), 533-558. https://doi.org/10.1007/s11336-016-9552-7

Maydeu-Olivares, A. i McArdle, J. J. (Eds.). (2005). *Contemporary psychometrics*. Psychology Press. https://doi.org/10.4324/9781410612977

McDonald, R. P. (2013). *Test theory: A unified treatment*. psychology press. https://doi.org/10.4324/9781410601087

McNeish, D., & Wolf, M. G. (2023). Dynamic fit index cutoffs for one-factor models. *Behavior Research Methods*, *55*(3), 1157-1174. https://doi.org/10.3758/s13428-022-01847-y

McNeish, D., An, J. i Hancock, G. R. (2018). The thorny relation between measurement quality and fit index cutoffs in latent variable models. *Journal of Personality Assessment*, *100*(1), 43-52. https://doi.org/10.1080/00223891.2017.1281286

Millsap, R. E. (2007). Structural equation modeling made difficult. *Personality and Individual Differences*, *42*(5), 875-881. https://doi.org/10.1016/j.paid.2006.09.021

Morin, A. J., Myers, N. D., & Lee, S. (2020). Modern factor analytic techniques: Bifactor models, exploratory structural equation modeling (ESEM), and bifactor-ESEM. *Handbook of Sport Psychology*, 1044-1073. https://doi.org/10.1002/9781119568124.ch51

Mur, J. A., Rivas, A., Trueba, D. A. i Girardi, C. I. P. (2022). The relationship between self-efficacy, state-trait anxiety and cognitive test anxiety: A study among university students in Argentina. *Psychological Thought*, *15*(2), 75. https://doi.org/10.37708/psyct.v15i2.664

Murray, A. L., Booth, T., Eisner, M., Obsuth, I., & Ribeaud, D. (2019). Quantifying the strength of general factors in psychopathology: A comparison of CFA with maximum likelihood estimation, BSEM, and ESEM/EFA bifactor approaches. *Journal of Personality Assessment*, *101*(6), 631-643. https://doi.org/10.1080/00223891.2018.1468338

Natesan, P. (2015). Comparing interval estimates for small sample ordinal CFA models. *Frontiers in Psychology*, *6*, 1599. https://doi.org/10.3389/fpsyg.2015.01599

Netemeyer, R. G., Bearden, W. O. i Sharma, S. (2003). *Scaling procedures: Issues and applications*. sage publications. https://doi.org/10.4135/9781412985772

Nguyen, L. H. i Holmes, S. (2019). Ten quick tips for effective dimensionality reduction. *PLoS Computational Biology*, *15*(6), e1006907. https://doi.org/10.1371/journal.pcbi.1006907

Niemand, T. i Mai, R. (2018). Flexible cutoff values for fit indices in the evaluation of structural equation models. *Journal of the Academy of Marketing Science*, *46*, 1148-1172. https://doi.org/10.1007/s11747-018-0602-9

Nilsen, F. A., Bang, H., Boe, O., Martinsen, Ø. L., Lang-Ree, O. C. i Røysamb, E. (2020). The Multidimensional Self-Control Scale (MSCS): Development and validation. *Psychological Assessment*, *32*(11), 1057. https://doi.org/10.1037/pas0000950

Norris, M. i Lecavalier, L. (2010). Evaluating the use of exploratory factor analysis in developmental disability psychological research. *Journal of Autism and Developmental Disorders*, *40*, 8-20. https://doi.org/10.1007/s10803-009-0816-2

Onyeizugbo, E. U. (2010). Self-efficacy, gender and trait anxiety as moderators of test anxiety. *Electronic Journal of Research in Educational Psychology*, *8*(1), 299-312. https://doi.org/10.25115/ejrep.v8i20.1377

Pan, J., Ip, E. H., & Dubé, L. (2017). An alternative to post hoc model modification in confirmatory factor analysis: The Bayesian lasso. *Psychological Methods*, *22*(4), 687. https://doi.org/10.1037/met0000112

Paulhus, D. L. i Vazire, S. (2007). The self-report method. in: *Handbook of research methods in personality psychology*, 1(2007), 224-239.

Peng, R. D. (2016). *R programming for data science* (pp. 86-181). Victoria, BC, Canada: Leanpub.

Pereira, T. L. i Cribari-Neto, F. (2014). Detecting model misspecification in inflated beta regressions. *Communications in Statistics-Simulation and Computation*, *43*(3), 631-656. https://doi.org/10.1080/03610918.2012.712183

Peterson, R. A. i Kim, Y. (2013). On the relationship between coefficient alpha and composite reliability. *Journal of Applied Psychology*, *98*(1), 194. https://doi.org/10.1037/a0030767

Pintea, S. i Moldovan, R. (2009). The receiver-operating characteristic (ROC) analysis: Fundamentals and applications in clinical psychology. *Journal of Evidence-Based Psychotherapies*, *9*(1), 49.

Preacher, K. J. i Merkle, E. C. (2012). The problem of model selection uncertainty in structural equation modeling. *Psychological Methods*, *17*(1), 1. https://doi.org/10.1037/a0026804

Price, L. R. (2016). *Psychometric methods: Theory into practice*. Guilford Publications.

Putnick, D. L. i Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71-90. https://doi.org/10.1016/j.dr.2016.06.004

R Core Team, (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: a comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, *87*(3), 517. https://doi.org/10.1037/0021-9010.87.3.517

Rao, C. R. i Sinharay, S. (Eds.). (2007). *Psychometrics* (Vol. 26, No. 26). elsevier.

Ray, P., Reddy, S. S. i Banerjee, T. (2021). Various dimension reduction techniques for high dimensional data analysis: a review. *Artificial Intelligence Review*, *54*(5), 3473-3515. https://doi.org/10.1007/s10462-020-09928-0

Raykov, T. i Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Routledge. https://doi.org/10.4324/9780203841624

Reichenheim, M. E., Moraes, C. L., Oliveira, A. S. i Lobato, G. (2011). Revisiting the dimensional structure of the Edinburgh Postnatal Depression Scale (EPDS): empirical evidence for a general factor. *BMC Medical Research Methodology*, *11*(1), 1-12. https://doi.org/10.1186/1471-2288-11-93

Reise, S. P., Kim, D. S., Mansolf, M. i Widaman, K. F. (2016). Is the bifactor model a better model or is it just better at modeling implausible responses? Application of iteratively reweighted least squares to the Rosenberg Self-Esteem Scale. *Multivariate Behavioral Research*, *51*(6), 818-838. https://doi.org/10.1080/00273171.2016.1243461

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*(3), 552. https://doi.org/10.1037/0033-2909.114.3.552

Rhemtulla, M. (2016). Population performance of SEM parceling strategies under measurement and structural model misspecification. *Psychological Methods*, *21*(3), 348. https://doi.org/10.1037/met0000072

Robins, R. W., Fraley, R. C. i Krueger, R. F. (Eds.). (2009). *Handbook of research methods in personality psychology*. Guilford Press.

Robitzsch, A. (2023). Modeling Model Misspecification in Structural Equation Models. *Stats*, *6*(2), 689-705. https://doi.org/10.3390/stats6020044

Rosenberg, M. (1965). *Society and the adolescent self-image* (Vol. 11, p. 326). Princeton, NJ: Princeton university press.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1-36. https://doi.org/10.18637/jss.v048.i02

Rytilä-Manninen, M., Fröjd, S., Haravuori, H., Lindberg, N., Marttunen, M., Kettunen, K. i Therman, S. (2016). Psychometric properties of the Symptom Checklist-90 in adolescent psychiatric inpatients and age-and gender-matched community youth. *Child and Adolescent Psychiatry and Mental Health*, *10*, 1-12. https://doi.org/10.1186/s13034-016-0111-x

S. Davcik, N. (2014). The use and misuse of structural equation modeling in management research: A review and critique. *Journal of Advances in Management Research*, *11*(1), 47-81. https://doi.org/10.1108/JAMR-07-2013-0043

Saris, W. E., Satorra, A. i Van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications?. *Structural Equation Modeling*, *16*(4), 561-582. https://doi.org/10.1080/10705510903203433

Schermelleh-Engel, K., Moosbrugger, H. i Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*(2), 23-74.

Schmitt, T. A., Sass, D. A., Chappelle, W. i Thompson, W. (2018). Selecting the "best" factor structure and moving measurement validation forward: An illustration. *Journal of Personality Assessment*, *100*(4), 345-362. https://doi.org/10.1080/00223891.2018.1449116

Schumacker, R. E., & Lomax, R. G. (2016). *A Beginner's Guide to Structural Equation Modeling* (wyd. 4). Routledge.

Sellbom, M. i Tellegen, A. (2019). Factor analysis in psychological assessment research: Common pitfalls and recommendations. *Psychological Assessment*, *31*(12), 1428. https://doi.org/10.1037/pas0000623

Shapiro, G. K., Tatar, O., Dube, E., Amsel, R., Knauper, B., Naz, A., ... i Rosberger, Z. (2018). The vaccine hesitancy scale: Psychometric properties and validation. *Vaccine*, *36*(5), 660-667. https://doi.org/10.1016/j.vaccine.2017.12.043

Sharma, S., Mukherjee, S., Kumar, A. i Dillon, W. R. (2005). A simulation study to investigate the use of cutoff values for assessing model fit in covariance structure models. *Journal of Business Research*, *58*(7), 935-943. https://doi.org/10.1016/j.jbusres.2003.10.007

Shek, D. T. i Yu, L. (2014). Confirmatory factor analysis using AMOS: A demonstration. *International Journal on Disability and Human Development*, *13*(2), 191-204. https://doi.org/10.1515/ijdhd-2014-0305

Shevlin, M. i Adamson, G. (2005). Alternative factor models and factorial invariance of the GHQ-12: a large sample analysis using confirmatory factor analysis. *Psychological assessment*, *17*(2), 231. https://doi.org/10.1037/1040-3590.17.2.231

Shi, D., Lee, T. i Maydeu-Olivares, A. (2019). Understanding the model size effect on SEM fit indices. *Educational and Psychological Measurement*, *79*(2), 310-334. https://doi.org/10.1177/0013164418783530

Shrestha, N. (2021). Factor analysis as a tool for survey analysis. *American Journal of Applied Mathematics and Statistics*, *9*(1), 4-11. https://doi.org/10.12691/ajams-9-1-2

Sideridis, G. D. i Jaffari, F. (2022). An R function to correct fit indices and omnibus tests in confirmatory factor analysis. *Measurement and Evaluation in Counseling and Development*, *55*(1), 48-70. https://doi.org/10.1080/07481756.2021.1906159

Sifuentes-Amaya, R. i Ramirez-Valverde, G. (2010). Effects of specifying an incorrect model for logistic regression, with two independent correlated variables. *Agrociencia*, *44*(2), 197-207.

Silva, R. B., Oliveira, D. D., Santos, D. P. D., Santos, L. F. D., Wilson, R. E. i Bêdo, M. V. N. (2020). Criteria for choosing the number of dimensions in a principal component analysis: An empirical assessment. *Anais*. https://doi.org/10.5753/sbbd.2020.13632

Singh, P., Singh, S. i Paprzycki, M. (2023). Detection and elimination of multicollinearity in regression analysis. *International Journal of Knowledge-based and Intelligent Engineering Systems*, (Preprint), 1-7. https://doi.org/10.3233/KES-221622

Sivo, S. A., Fan, X., Witta, E. L. i Willse, J. T. (2006). The search for" optimal" cutoff properties: Fit index criteria in structural equation modeling. *The Journal of Experimental Education*, *74*(3), 267-288. https://doi.org/10.3200/JEXE.74.3.267-288

Slocum-Gori, S. L. i Zumbo, B. D. (2011). Assessing the unidimensionality of psychological scales: Using multiple criteria from factor analysis. *Social Indicators Research*, *102*, 443-461. https://doi.org/10.1007/s11205-010-9682-8

Song, X. Y. i Lee, S. Y. (2005). Maximum likelihood analysis of nonlinear structural equation models with dichotomous variables. *Multivariate Behavioral Research*, *40*(2), 151-177. https://doi.org/10.1207/s15327906mbr4002_1

Sosnowski T. i Wrześniewski K. (1983) Inwentarz do Pomiaru Stanu i Cechy Lęku /STAI/ C. D. Spielbergera i jego polska adaptacja. [w:] Wybrane zagadnienia lęku. Teoria i pomiar. Wrześniewski K (red). Warszawska Akademia Medyczna, Warszawa: 138-153.

Stanley, K. D. i Murphy, M. R. (1997). A comparison of general self-efficacy with self-esteem. *Genetic, Social, and General Psychology Monographs*, *123*(1), 79-100.

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences*, *42*(5), 893-898. https://doi.org/10.1016/j.paid.2006.09.017

Stone, B. M. (2021). The ethical use of fit indices in structural equation modeling: Recommendations for psychologists. *Frontiers in Psychology*, *12*, 783226. https://doi.org/10.3389/fpsyg.2021.783226

Strelhow, M. R. W., Sarriera, J. C. i Casas, F. (2020). Evaluation of well-being in adolescence: Proposal of an integrative model with hedonic and eudemonic aspects. *Child Indicators Research*, *13*, 1439-1452. https://doi.org/10.1007/s12187-019-09708-5

Sun, S., Konold, T. R. i Fan, X. (2011). Effects of latent variable nonnormality and model misspecification on testing structural equation modeling interactions. *The Journal of Experimental Education*, *79*(3), 231-256. https://doi.org/10.1080/00220973.2010.481683

Taasoobshirazi, G. i Wang, S. (2016). The performance of the SRMR, RMSEA, CFI, and TLI: An examination of sample size, path size, and degrees of freedom. *Journal of Applied Quantitative Methods*, *11*(3), 31-39.

Taherdoost, H. (2019). What is the best response scale for survey and questionnaire design; review of different lengths of rating scale/attitude scale/Likert scale. *Hamed Taherdoost*, 1-10.

Tanujaya, B., Prahmana, R. C. I. i Mumu, J. (2022). Likert scale in social sciences research: Problems and difficulties. *Journal of Social Sciences*, 2022, Vol 16, No.4, 89-101 https://doi.org/10.51709/19951272/Winter2022/7

Tay, R. (2017). Correlation, variance inflation and multicollinearity in regression model. *Journal of the Eastern Asia Society for Transportation Studies*, *12*, 2006-2015.

Thompson, B. (2004). Exploratory and confirmatory factor analysis: Understanding concepts and applications. *Washington, DC*, *10694*(000), 3. https://doi.org/10.1037/10694-000

Thurber, S. i Bonynge, M. (2011). SEM-based composite reliability estimates of the Crisis Acuity Rating Scale with children and adolescents. *Archives of Assessment Psychology*, *1*(1), 1-9.

Timmerman, M. E., Lorenzo-Seva, U. i Ceulemans, E. (2018). The number of factors problem. *The Wiley handbook of psychometric testing: a multidisciplinary reference on survey, scale and test development*, 305-324. https://doi.org/10.1002/9781118489772.ch11

Uanhoro, J. (2017). Misspecification and fit indices in covariance-based SEM. Retrieved November, 2023, from https://www.jamesuanhoro.com/post/2017/10/28/misspecification-and-fit-indices-in-covariance-based-sem/

Unal, I. (2017). Defining an optimal cut-point value in ROC analysis: an alternative approach. *Computational and mathematical methods in medicine*, *2017*. https://doi.org/10.1155/2017/3762651

Urbina, S. (2014). *Essentials of psychological testing*. John Wiley i Sons. https://doi.org/10.1002/9781394259458

Ventura-León, J., Peña-Calero, B. N. i Burga-León, A. (2023). The effect of normality and outliers on bivariate correlation coefficients in psychology: A Monte Carlo simulation. *The Journal of General Psychology*, *150*(4), 405-422. https://doi.org/10.1080/00221309.2022.2094310

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological methods*, *17*(2), 228. https://doi.org/10.1037/a0027127

Wang, J. i Wang, X. (2019). *Structural equation modeling: Applications using Mplus*. John Wiley i Sons. https://doi.org/10.1002/9781119422730

Wang, J., Yue, S., Yu, X. i Wang, Y. (2017). An efficient data reduction method and its application to cluster analysis. *Neurocomputing*, *238*, 234-244. https://doi.org/10.1016/j.neucom.2017.01.059

Wang, K., Xu, Y., Wang, C., Tan, M. i Chen, P. (2020). A Corrected Goodness-of-Fit Index (CGFI) for model evaluation in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(5), 735-749. https://doi.org/10.1080/10705511.2019.1695213

Wang, Y. A. i Rhemtulla, M. (2021). Power analysis for parameter estimation in structural equation modeling: A discussion and tutorial. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920918253. https://doi.org/10.1177/2515245920918253

Werner, C. i Schermelleh-Engel, K. (2010). Deciding between competing models: Chi-square difference tests. *Introduction to structural equation modeling with LISREL*, 1-3.

West, S. G., Taylor, A. B. i Wu, W. (2012). Model fit and model selection in structural equation modeling. *Handbook of structural equation modeling*, *1*, 209-231.

West, S. G., Wu, W., McNeish, D. i Savord, A. (2023). Model fit in structural equation modeling. *Handbook of structural equation modeling*, 184-205.

Westland, J. C. (2015). Structural equation models. *Studies in Systems, Decision and Control*, *22*(5), 152. https://doi.org/10.1007/978-3-319-16507-3

Whittaker, T. A. (2012). Using the modification index and standardized expected parameter change for model modification. *The Journal of Experimental Education*, *80*(1), 26-44. https://doi.org/10.1080/00220973.2010.531299

Winship, C. i Western, B. (2016). Multicollinearity and model misspecification. *Sociological Science*, *3*(27), 627-649. https://doi.org/10.15195/v3.a27

Wolf, M. G. i McNeish, D. (2023). dynamic: An R package for deriving dynamic fit index cutoffs for factor analysis. *Multivariate Behavioral Research*, *58*(1), 189-194. https://doi.org/10.1080/00273171.2022.2163476

Ximénez, C., Maydeu-Olivares, A., Shi, D. i Revuelta, J. (2022). Assessing cutoff values of SEM fit indices: Advantages of the unbiased SRMR index and its cutoff criterion based on communality. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(3), 368-380. https://doi.org/10.1080/10705511.2021.1992596

Yang, B. (2005). Factor analysis methods. In: *Research in organizations: Foundations and methods of inquiry*, (181-199).

Yang, Y. i Green, S. B. (2010). A note on structural equation modeling estimates of reliability. *Structural Equation Modeling*, *17*(1), 66-81. https://doi.org/10.1080/10705510903438963

Yu, S. C., Lin, Y. H. i Hsu, W. H. (2013). Applying structural equation modeling to report psychometric properties of Chinese version 10-item CES-D depression scale. *Quality i Quantity*, *47*, 1511-1518. https://doi.org/10.1007/s11135-011-9604-0

Yuan, K. H., Fung, W. K. i Reise, S. P. (2004). Three Mahalanobis distances and their role in assessing unidimensionality. *British Journal of Mathematical and Statistical Psychology*, *57*(1), 151-165. https://doi.org/10.1348/000711004849231

Yuan, K. H. i Bentler, P. M. (2007). Robust procedures in structural equation modeling. In *Handbook of latent variable and related models* (pp. 367-397). North-Holland. https://doi.org/10.1016/B978-044452044-9/50020-3

Zheng, B. Q. i Bentler, P. M. (2024). Enhancing Model Fit Evaluation in SEM: Practical Tips for Optimizing Chi-Square Tests. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-6. https://doi.org/10.1080/10705511.2024.2354802

Ziegler, M. i Hagemann, D. (2015). Testing the unidimensionality of items. *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000309

SUPLEMENT: DODATKOWE TABELE I ZAŁĄCZNIKI

Tabela 10
Statystyki opisowe wskaźników dopasowania w badaniu symulacyjnym (*N* = 31500)

	R	M	SD	Mdn	Sk	Kurt	D
χ^2	7.23-9099.90	603.57	545.59	445.06	1.33	4.71	.14**
χ^2 / df	.29-7.76	1.19	.33	1.11	3.10	18.12	.18**
χ^2_p	.00-1.00	.22	.27	.09	1.18	.21	.21**
CFI	.25-1.00	.98	.06	.99	-5.37	38.25	.34**
TLI	-9.81-6.53	.98	.12	.99	-27.92	3118.11	.33**
GFI	.41-1.00	.93	.07	.95	-2.08	4.86	.18**
AGFI	.35-1.00	.91	.08	.94	-2.04	4.54	.18**
RNI	-8.33-5.95	.98	.10	.99	-27.22	3295.00	.32**
RMSEA	.00-.26	.02	.02	.01	2.11	8.32	.16**
AIC	-87973.01-196616.13	28135.40	28992.55	19779.25	1.54	3.56	.13**
BIC	-87482.24-197155.98	28398.80	29068.90	20005.06	1.54	3.54	.13**
RMR	.00-.28	.02	.03	.01	3.28	15.15	.21**
SRMR	.00-.10	.03	.02	.02	1.36	2.28	.10**
χ^2_{diff}	-8.26-72.47	4.75	5.71	2.79	1.90	6.20	.20**
$\chi^2_{diff\text{-}p}$	.00-1.00	.55	.35	.55	-.07	-1.44	.11**
nm%	.00-100.00	63.26	34.51	79.52	-.98	-.66	.21**
m%	.00-100.00	3.59	7.71	1.05	6.90	68.07	.32**
EPC nm%	.00-100.00	8.67	12.77	5.36	3.64	15.61	.27**
EPC m%	.00-100.00	1.43	8.03	.00	8.31	77.68	.43**
i%	.00-100.00	23.05	33.22	8.00	1.43	.38	.27**
total m%	.00-100.00	5.02	11.67	1.43	5.50	36.14	.33**
mean MI	.35-350637.61	102.07	3735.30	1.19	69.49	5432.45	.49**
mean EPC	-73.73-740.27	.10	5.26	.00	110.39	13958.06	.45**
mean NCP	.08-121707.43	696.60	5179.55	42.01	13.80	219.56	.45**

* p <.05,** p <.01

Tabela 11
Statystyki opisowe wskaźników dopasowania w badaniu empirycznym (*N* = 15000)

	R	M	SD	Mdn	Sk	Kurt	D
χ^2	7.23-9099.90	603.57	545.59	445.06	1.33	4.71	.14**
χ^2 / df	.29-7.76	1.19	.33	1.11	3.10	18.12	.18**
χ^2_p	.00-1.00	.22	.27	.09	1.18	.21	.21**
CFI	.25-1.00	.98	.06	.99	-5.37	38.25	.34**
TLI	-9.81-6.53	.98	.12	.99	-27.92	3118.11	.33**
GFI	.41-1.00	.93	.07	.95	-2.08	4.86	.18**
AGFI	.35-1.00	.91	.08	.94	-2.04	4.54	.18**
RNI	-8.33-5.95	.98	.10	.99	-27.22	3295.00	.32**
RMSEA	.00-.26	.02	.02	.01	2.11	8.32	.16**
AIC	-87973.01-196616.13	28135.40	28992.55	19779.25	1.54	3.56	.13**
BIC	-87482.24-197155.98	28398.80	29068.90	20005.06	1.54	3.54	.13**
RMR	.00-.28	.02	.03	.01	3.28	15.15	.21**
SRMR	.00-.10	.03	.02	.02	1.36	2.28	.10**
χ^2_{diff}	-8.26-72.47	4.75	5.71	2.79	1.90	6.20	.20**
$\chi^2_{diff\text{-}p}$	.00-1.00	.55	.35	.55	-.07	-1.44	.11**
nm%	.00-100.00	63.26	34.51	79.52	-.98	-.66	.21**
m%	.00-100.00	3.59	7.71	1.05	6.90	68.07	.32**
EPC nm%	.00-100.00	8.67	12.77	5.36	3.64	15.61	.27**
EPC m%	.00-100.00	1.43	8.03	.00	8.31	77.68	.43**
i%	.00-100.00	23.05	33.22	8.00	1.43	.38	.27**
total m%	.00-100.00	5.02	11.67	1.43	5.50	36.14	.33**
mean MI	.35-350637.61	102.07	3735.30	1.19	69.49	5432.45	.49**
mean EPC	-73.73-740.27	.10	5.26	.00	110.39	13958.06	.45**
mean NCP	.08-121707.43	696.60	5179.55	42.01	13.80	219.56	.45**

* $p < .05$, ** $p < .01$

Tabela 12
Związki pomiędzy poszczególnymi indeksami dopasowania w badaniu symulacyjnym

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	χ^2															
2	χ^2 / df	.33**														
3	χ^2_p	-.62**	-.91**													
4	CFI	-.49**	-.71**	.77**												
5	TLI	-.46**	-.72**	.76**	.99**											
6	GFI	-.74**	-.38**	.59**	.66**	.64**										
7	AGFI	-.66**	-.40**	.57**	.67**	.65**	.99**									
8	RNI	-.49**	-.71**	.77**	.99**	.99**	.66**	.67**								
9	RMSEA	.32**	.93**	-.86**	-.82**	-.82**	-.56**	-.60**	-.81**							
10	RMR	.01*	-.20**	.16**	-.10**	-.09**	-.23**	-.24**	-.10**	-.04**						
11	SRMR	.13**	-.04**	-.01	-.51**	-.49**	-.46**	-.49**	-.49**	.16**	.44**					
12	AIC	.35**	-.29**	.13**	.20**	.22**	.11**	.20**	.20**	-.42**	.33**	-.14**				
13	BIC	.36**	-.29**	.13**	.20**	.22**	.11**	.20**	.19**	-.42**	.33**	-.14**	99**			
14	χ^2_{diff}	-.01	.05**	-.04**	-.02**	-.03**	-.00	-.03**	-.02**	.05**	-.04**	-.03**	-.04**	-.03**		
15	$\chi^2_{diff\text{-}p}$	-.01*	-.08**	.07**	.05**	.05**	.03**	.04**	.05**	-.07**	.05**	.02**	.05**	.05**	-.82**	

* p <.05,** p <.01

Tabela 13
Związki pomiędzy poszczególnymi indeksami dopasowania w badaniu empirycznym

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	χ^2															
2	χ^2 / df	.71**														
3	χ^2_p	-.61**	-.57**													
4	CFI	-.58**	-.66**	.47**												
5	TLI	-.55**	-.68**	.45**	.98**											
6	GFI	-.54**	-.47**	.41**	.95**	.90**										
7	AGFI	-.50**	-.60**	.38**	.96**	.98**	.93**									
8	RNI	-.58**	-.66**	.47**	.99**	.99**	.99**	.99**								
9	RMSEA	.48**	.65**	-.38**	-.95**	-.98**	-.89**	-.97**	-.95**							
10	RMR	.54**	.64**	-.41**	-.97**	-.97**	-.93**	-.97**	-.97**	.95**						
11	SRMR	.49**	.61**	-.36**	-.97**	-.97**	-.92**	-.97**	-.97**	.96**	.99**					
12	AIC	.79**	.48**	-.50**	-.02	-.00	.04**	.08**	-.02	-.09**	-.03**	-.08**				
13	BIC	.79**	.47**	-.50**	-.01	-.00	.04**	.08**	-.01	-.09**	-.03**	-.08**				
14	χ^2_{diff}	.23**	.36**	-.19**	-.09**	-.21**	.03**	-.18**	-.09**	.18**	.17**	.17**	.24**	.24**		
15	$\chi^2_{diff\text{-}p}$	-.28**	-.41**	.24**	.14**	.24**	.02*	.20**	.14**	-.21**	-.21**	-.20**	-.28**	-.28**	-.98**	

* p <.05,** p <.01

Tabela 14
Związki pomiędzy poszczególnymi indeksami modyfikacji w badaniu symulacyjnym

	1	2	3	4	5	6	7	8	9
1 nm%									
2 m%	-.77**								
3 EPC nm%	.27**	-.27**							
4 EPC m%	-.22**	.14**	.34**						
5 i%	-.74**	.65**	-.63**	-.20**					
6 total m%	-.81**	.93**	-.14**	.37**	.52**				
7 mean MI	-.32**	.40**	.58**	.42**	-.16**	.52**			
8 mean EPC	-.20**	.12**	.02**	.09**	.15**	.13**	.09**		
9 mean NCP	.49**	-.45**	.66**	.02**	-.62**	-.40**	.24**	-.08**	

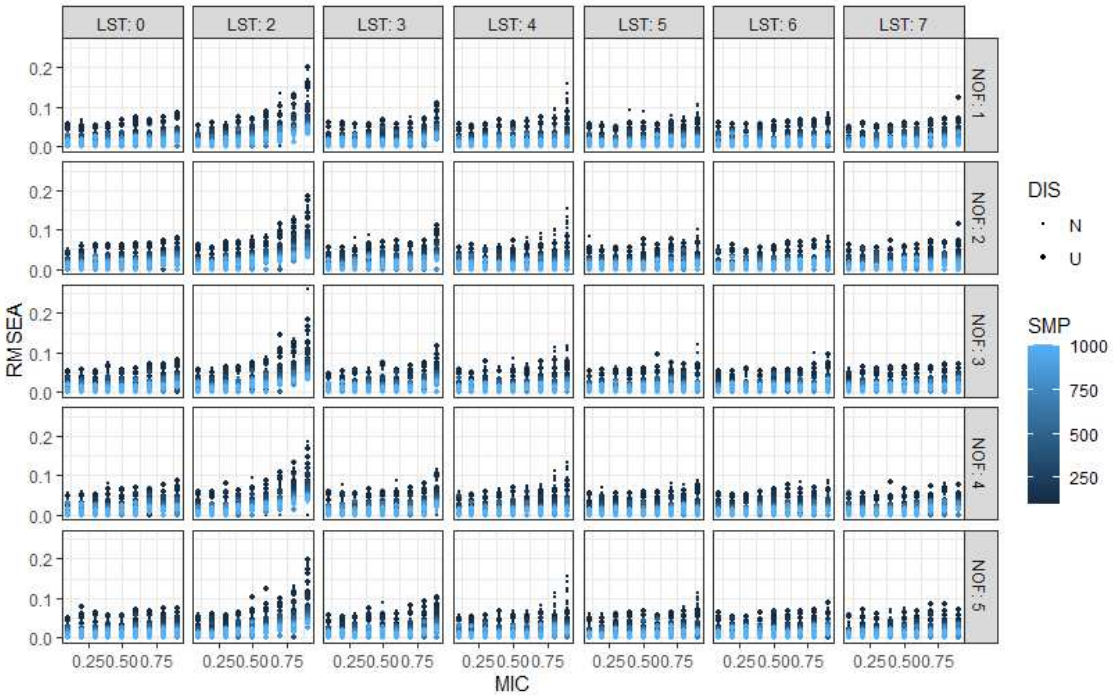
* p <.05,** p <.01

Tabela 15
Związki pomiędzy poszczególnymi indeksami modyfikacji w badaniu empirycznym

	1	2	3	4	5	6	7	8	9
1 nm%									
2 m%	-.50**								
3 EPC nm%	.05**	-.58**							
4 EPC m%	-.53**	-.00	.30**						
5 i%	-.16**	.62**	-.79**	-.36**					
6 total m%	-.83**	.56**	-.14**	.75**	.07**				
7 mean MI	-.50**	-.05**	.57**	.82**	-.57**	.64**			
8 mean EPC	-.22**	.29**	-.17**	.18**	.19**	.32**	.16**		
9 mean NCP	.36**	-.51**	.70**	-.12**	-.66**	-.39**	.26**	-.06**	

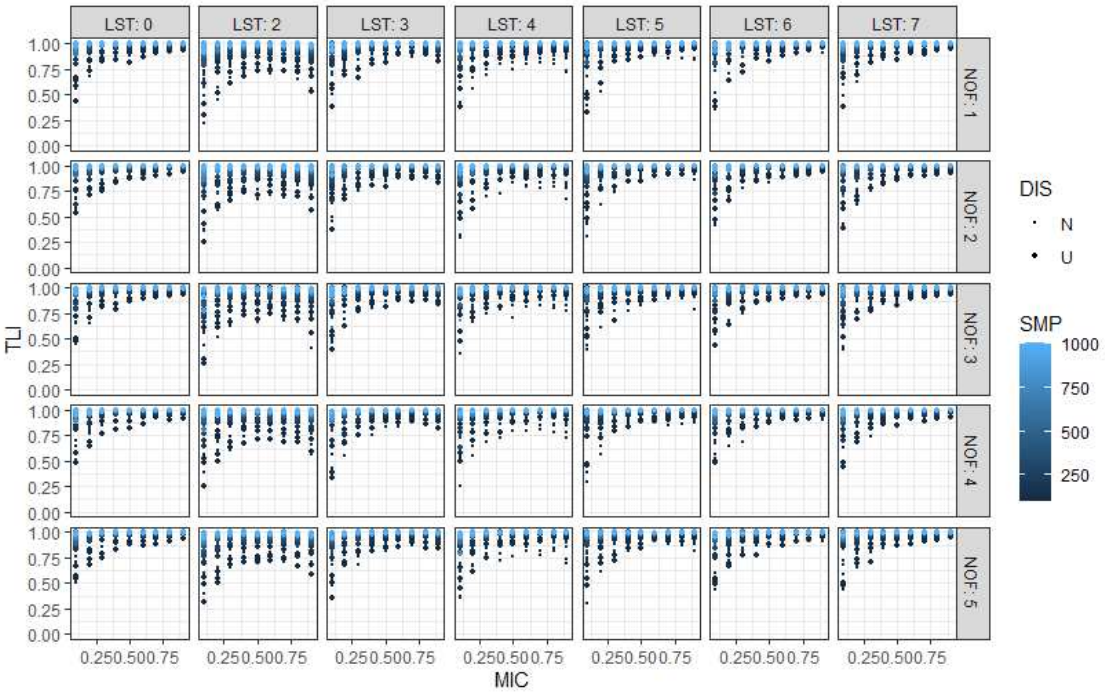
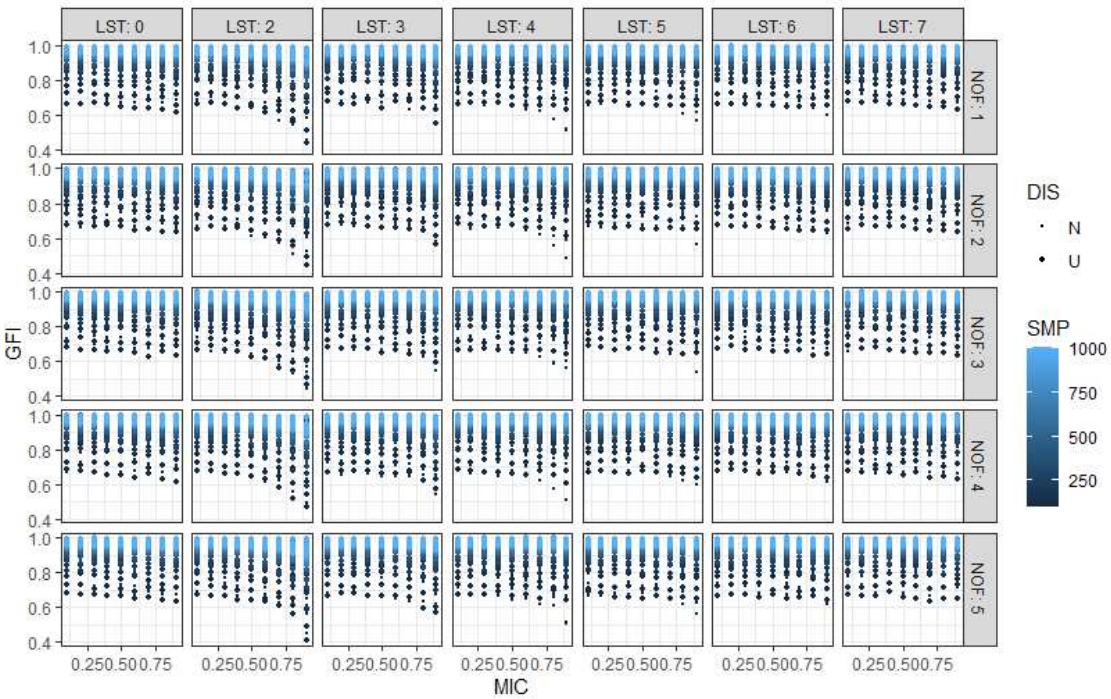
* p <.05,** p <.01

Rysunek 55. Poziom RMSEA dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST), oraz różnej liczby czynników (NOF) i wielkości próby (SMP)



Rysunek 56. Poziom χ2 dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST), oraz różnej liczby czynników (NOF) i wielkości próby (SMP)

Rysunek 57. Poziom TLI dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST), oraz różnej liczby czynników (NOF) i wielkości próby (SMP)



Rysunek 58. Poziom GFI dla poszczególnych modeli w zależności od korelacji między pozycjami (MIC), różnych typów skal (LST) i rozkładów (DIST), oraz różnej liczby czynników (NOF) i wielkości próby (SMP)